Design Utility Methods for Preferentially-Sampled Spatial Data

Gray, Elizabeth

Award date:
2021

Awarding institution:
University of Bath

Design Utility Methods for Preferentially-Sampled Spatial Data

Gray, Elizabeth

Award date:
2021

Design Utility Methods for Preferentially-Sampled Spatial Data

Elizabeth Gray

Thesis submitted for doctoral degree

University of Bath

Department of Mathematical Sciences

August 2021

The material presented here for examination for the award of a higher degree by research has not been incorporated into a submission for another degree.

I am the author of this thesis, and the work described therein was carried out by myself personally.

Acknowledgements

I would like to give special thanks to my supervisor, Evangelos Evangelou, for all his excellent guidance and continual encouragement over the last four years. I am also very grateful for everyone from SAMBa, and all my fantastic friends in both the Maths department, and elsewhere in Bath, for helping to make doing a PhD such an enjoyable experience.

Abstract

Spatial preferential sampling occurs when the choice of sampling locations at which a spatial process of interest is measured is stochastically dependent on the values of this process. Ignoring such a sampling scheme when mapping the spatial process leads to inaccurate predictions, particularly at locations further from the sampling sites. This may be dealt with by jointly modelling the sampling process with the spatial process. Existing methods for this require that the sampling locations be independent of one another. In this thesis we dispense with such an unrealistic requirement and model the sampling process as a whole. This is achieved by defining a whole-design utility function over the space of possible sampling designs to assign a 'usefulness' to each design based on a set of possible experimenter preferences and some (unknown) strength of preference parameters. We may then assume that the probability that any design is selected is proportional to this utility. We shall give particular attention to utility functions which balance a preference for high values with a preference for even coverage of the region of interest. This whole-design approach presents a variety of challenges which we will address, such as the definition of suitable utility functions and the selection of suitable algorithms for the fitting of such models, in particular, because the the design distribution may introduce intractable normalising constants into the likelihood function of the spatial process. We shall describe methods for replacing these with suitable approximations, explore methods for efficiently drawing samples of designs required to form these approximations, and define a class of utility functions, the properties of which allow for the application of a combinatorially-based method to improve accuracy. We shall demonstrate the effectiveness of our model via a simulation study and application to various spatial data sets. Finally, we shall consider methods of combining multiple data sets in order to get better estimates of strength of preference.

Contents

Chapter 1

Introduction and motivation

1.1 Preferential sampling and its effects

Preferential sampling refers to the situation in which the choice of sampling units is stochastically dependent on the value of the quantity to be measured. Spatially, this means that the choice of sampling locations, the sampling design, would be dependent on the values of the spatial process of interest, a quantity such as air or soil pollution. Preferential sampling may take place for several reasons. Where the process of interest is air pollution, the investigators (with a limited budget and a finite possible sampling network) may be interested in identifying where pollution guidelines are exceeded, say close to industrial areas or main roads. Similarly, where investigators seek to identify a trend between two variables, it would make sense to sample in particularly high or low areas. Preferential sampling presents a problem because if we only have access to the high-value measurements we will overestimate the low-valued areas.

The effects of preferential sampling have been demonstrated by many authors. Shaddick and Zidek (2014) consider the UK Smoke and Sulphur Dioxide Network, which monitored air pollution levels in the UK between the early 1960s and 2006. As UK air pollution declined and became of less governmental interest the network was gradually shut down from 1200 sites in 1971, to 65 in 2006. A predisposition for closing sites measuring lower values was shown, as was a resulting overestimation in air pollution levels. Fernández et al. (2005) compare the effects of different sampling designs in the monitoring of heavy metal content in moss, concluding that, in terms of finding the certain descriptive statistics, preferential sampling of high-valued regions leads to inaccurate inferences. Similarly, Gelfand et al. (2012) show, via a simulation study and making 'expected comparisons' of predicted surfaces, that preferential sampling based on some intensity, for example, a population density surface, has an impact on predictions when compared with sampling under complete spatial randomness. They show that using a non-preferential model does not sufficiently account for preference in the design, even when the preferential intensity is included as a predictor for the spatial process. Grisotto et al. (2016), again in the context of air pollution monitoring, demonstrate how ignoring preferential sampling when constructing prediction surfaces leads to a mis-specification of the uncertainty in the surface for more sparsely covered regions. Lee et al. (2015) investigate the knock-on effects of preferential sampling with respect to health effect analysis in an air pollution context: ordinary kriging is used to obtain estimates of air pollution

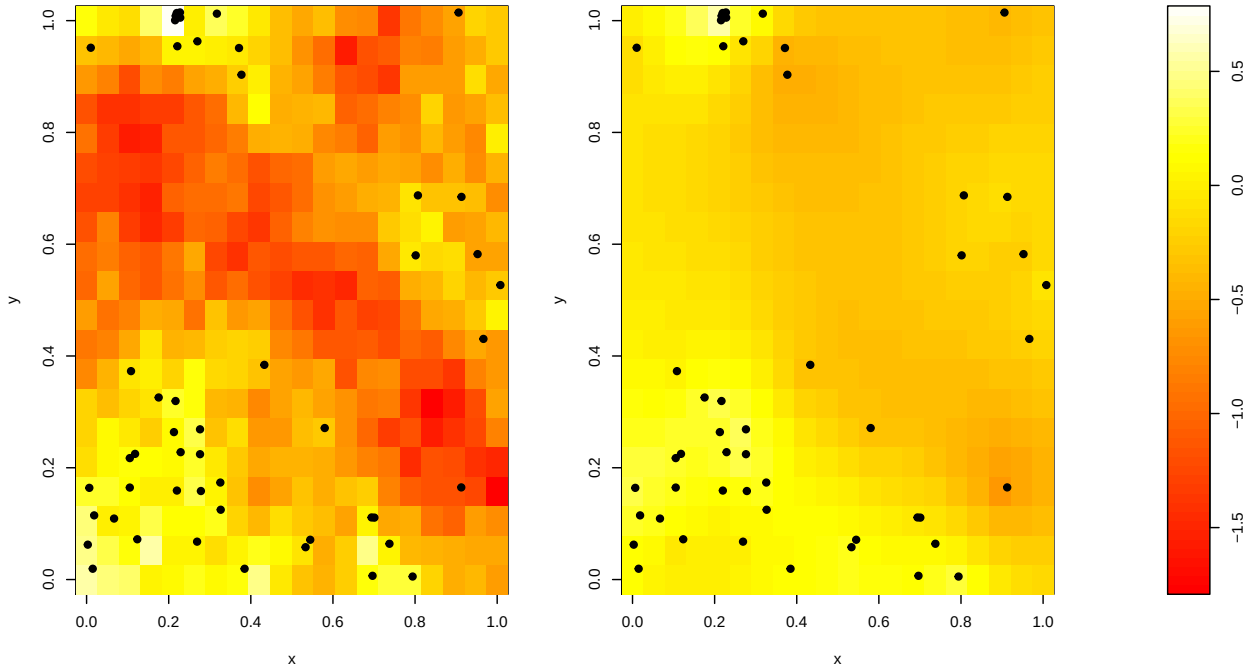


Figure 1.1: Two realisations of Gaussian random fields: on the left we have the original field with sampling locations superimposed. On the right we have the predicted field, reconstructed from 'measurements' with Gaussian measurement error taken at the sampling locations.

at sites at which health data is available. From these the health effect of air pollution is estimated. They demonstrate, via simulation study, that the validity of health effect predictions can be affected, with different scientific conclusions reached due to preference in the sampling design. Conn et al. (2017) discuss, in the context of animal population monitoring, how covariate parameters may be overestimated in the situation where volunteer data collectors with access to possible habitats may visit sites where they think it likely animals will be more often. Likewise they demonstrate that ignoring preferential sampling can lead to up to a forty percent bias in estimations of population density. Michalcová et al. (2011) conclude, via a comparison with data sets collected using stratified sampling, that preferential sampling has led to biased plots within vegetation databases.

We demonstrate the possible consequences of spatial preferential sampling on surface prediction in Figure 1.1, in which we show a predicted Gaussian random field reconstructed from a preferentially-sampled set of 'measurements'. Clearly, the reconstruction has missed the fact that there should be a lower-valued region in the central, lesser-sampled area.

1.1.1 Traditional Spatial Prediction: Gaussian random fields

The continuous spatial surface of interest may, in general, be assumed to be a Gaussian random field. We begin by briefly giving some background to what this is and how it may be used for spatial prediction (without necessarily taking account of the way in which the choice of sampling locations has been made). More detailed treatment of these concepts may be found in (Cressie, 2015) and (Gelfand et al., 2010).

A Gaussian random field $\{Z(s) : s \in S\}$, which takes values at locations s in some continuous spatial domain S, is such that every finite set of these values is a realisation of a multivariate normal distribution. These multivariate normal values have mean $\mathrm{E}(Z(s)) = \mu(s)$ (which may depend on spatially-varying covariates) and covariances $\mathrm{Cov}(Z(s_i), Z(s_j)) = \sigma^2 C(s_i, s_j)$, where σ^2 is a variance parameter, and C a correlation function.

A Gaussian random field is said to be stationary if the distribution of its values is invariant under constant spatial shifts. In other words, the mean μ is a constant over space, and the correlation function depends only on the difference between the points in question, rather their absolute locations in space. If we restrict the definition further, such that the correlation function is a function only of the absolute Euclidean distance (rather than directioned difference), then the corresponding field is said to be isotropic.

Popular choices for correlation functions displaying this property are the Matern class of functions, which include the exponential correlation function: $C_{\exp}(s_i, s_j) = \exp\left(|s_i - s_j|/\varphi\right)$, and the Gaussian correlation function: $C_{\mathrm{Gaus}}(s_i, s_j) = \exp\left(|s_i - s_j|^2/\varphi\right)$ with correlation parameter φ.

As the values of the Gaussian random field are multivariate normal, given n observed values $Z^{(1)} = \{Z(s_1), \ldots, Z(s_n)\}$ at locations with mean values $\mu^{(1)} = \{\mu(s_1), \ldots, \mu_(s_n)\}$ the values of the process at the m locations at which we would like to make predictions $Z^{(2)} = \{Z(s_1^{'}), \ldots, Z(s_m^{'})\}$ at locations with mean values $\mu^{(2)} = \{\mu(s_1^{'}), \ldots, \mu(s_m^{'})\}$ will follow a conditional multivariate normal distribution, with mean $\hat{\mu}$ and covariance $\sigma^2\hat{C}$ where $\hat{\mu} = \mu^2 + C_{21}C_{11}^{-1}(Z^1 - \mu^1)$ and $\hat{C} = C_{22} - C_{21}C_{11}^{-1}C_{12}$ where $C_{11}, C_{12}, C_{21}, C_{22}$ are the correlation matrices for the elements of Z^1 and Z^2 (both within: C_{11}, C_{22} and between: C_{12}, C_{21}). Taking this conditional mean as our prediction in such a way is known as Gaussian process regression (as the predicted values are weighted linear combinations of the observed values) or Kriging, (after Danie G. Krige who applied techniques of weighting measurements by distances to estimate subterranean gold deposits from boreholes (Krige, 1951)).

1.1.2 Poisson processes and Log-Gaussian Cox processes

A spatial point process may be used to describe the distribution of a pattern of points over a spatial domain S. The simplest of these is the Poisson process. This is defined such that, for a bounded spatial region $A \subset S$, and a function $\{\lambda(s) : s \in S,\ \lambda(s) \in \mathbb{N}\}$, the numbers of points in A follows a Poisson distribution with rate $\int_A \lambda(s)ds$. This function $\lambda(s)$ is known as the intensity function. Additionally, for another bounded spatial region $B \subset S$ which is disjoint with A, the number of points within A and B are independent of one another. Such a Poisson process is said to be homogeneous if λ is a constant over space, and inhomogeneous otherwise. A log-Gaussian Cox process (LGCP) (Møller et al., 1998) is a special case of an inhomogeneous Poisson process, such that log of the intensity function is itself a Gaussian random field, defined over the spatial domain S. This property makes LGCPs particularly attractive to the modelling of how a continuous spatial process (which may be treated as a Gaussian random field) has been sampled.

1.2 Existing approaches for preferentially-sampled data

Existing approaches to preferential sampling fall broadly into two categories: firstly, those that model the sampling process as a point process and are predominantly focused on building a map of the underlying process, and those that allocate selection probabilities to sites to inform weights, which are more focused on estimation of regression parameters for covariates upon which the preferentially-sampled observed response variables depend.

1.2.1 Designs as point processes

We first consider the approaches that treat the design as a point process. Considering the design as a random variable enables us to encapsulate the uncertainty of the experimenter with respect to the underlying field, and the fact that there is naturally an element of randomness in the choice of design where different designs may be equally useful to the experimenter.

Diggle et al. (2010): the sampling process as an LGCP.

Diggle et al. (2010) introduce the concept of preferential sampling in the context of geostatistical modelling. Where Z is the surface of interest, assumed to be a Gaussian random field, D the set of sampling locations, and Y the measurements taken at them, the authors define preferential sampling to be the case in which we have

$$P(Z, D) \neq P(Z)P(D).$$

They propose a model in which the sampling process should be assumed to be a log-Gaussian Cox process (LGCP), with intensity dependent on the value of the underlying surface Z, at location s via the intensity function

$$\lambda(s) = \exp(\alpha + \beta Z(s)),$$

where a high, positive β would indicate high levels of preferential sampling, $\beta < 0$ a tendency to favour low values of Z, and $\beta = 0$, the absence of preferential sampling. Conditional on the values of Z at their respective locations, the measurements $Y = \{y_1, \ldots, y_n\}$ are then assumed to be i.i.d normal with mean Z:

$$y_i | Z(s_i) \sim N(Z(s_i), \tau^2), \qquad i = 1, \ldots, n.$$

This shared process model may be fitted via maximum likelihood estimation with Monte Carlo approximations.

Pati et al (2011): shared predictors for the spatial and design process.

Pati et al. (2011) follow a similar route, in that the sampling process is modelled as a log Gaussian Cox process (LGCP). However, instead of including the value of the underlying spatial surface $Z(s)$ in the predictor of the intensity, they assume that it is the sum of two spatial processes: $\gamma(s)$ which is included in both the sampling intensity and the predictor of the measurements, and $\eta(s)$, representing spatial

random effects that are not included in the site selection process:

$$y_i \sim N(\eta(s_i) + \gamma(s_i), \sigma^2),$$

$$p(s) \ \propto \ \beta \exp(\gamma(s)),$$

where $p(s)$ denotes the intensity of the process by which the sampling process is modelled. The parameter β here is again a measure of how preferential the sampling is. This approach has the flexibility of allowing the sampling process to depend on covariates shared with the spatial surface without necessarily depending on the surface as a whole.

da Silva Ferreira (2020): preferential sampling with repulsion windows

da Silva Ferreira (2020) presents a model in which some level of repulsion between monitors may be accounted for. Here, the assumption is made that there exists a repulsion window of a fixed size around each sampling location, in which the experimenter is unwilling to situate another monitor, despite their intention to sample preferentially. This is achieved via the partition of the spatial region into regular squares with the largest possible area such that there is at most one monitor in each. Each of these partition squares is then associated with a value z_i relating to the continuous spatial process within it, i.e. an integral over its area. The sampling process is then treated as a Bernoulli process, with the probability that square i, associated with z_i, contains a monitor given by

$$p_i = 1 - \exp\left(-\Delta \exp(\alpha + \beta z_i)\right),$$

using the notation of Diggle et al. (2010) and Δ the side length of the partition squares. While this model goes some way towards accounting for a mixed preference for higher values, and a desire to space monitors out, there are some limitations.

The (fixed) level of repulsion is defined solely by the tightest spatial cluster: all windows, even those in areas of low point density must be small enough that in the high point density area, no window would contain more than one monitor. Any substantial attempts to spread monitors more evenly elsewhere must be ignored because the preference for spreading monitors out has been overwhelmed by high-value preference in one particular area. This situation is a very realistic possibility, as in the case in which an experimenter had a general preference for even coverage, in addition to some special interest in a specific area. This method carries the implicit assumption that the experimenter has no spacing related preferences beyond the scope of the repulsion windows, and, if they had had no high-value preferences, would have been just as likely to choose a design with all monitors in neighbouring 'windows', as one with monitors spread evenly across the region.

1.2.2 Inverse weighting approach

Scott and Wild (2011): Horwitz Thompson adjustment and Conditional Maximum Likelihood.

Scott and Wild (2011) address the problem of carrying out regression, with parameters β to be esti-

mated, on preferentially-sampled data $y_1, \ldots, y_n$ with covariates $\mathbf{x}_1, \ldots, \mathbf{x}_n$, from a population of size N. They describe several methods, firstly involving the Horwitz-Thompson weighting method (Horvitz and Thompson, 1952). Ordinarily, the parameters would be estimated by solving

$$\sum_i \frac{\partial}{\partial \beta} \log(f(y_i|\beta, \mathbf{x}_i)) = 0. \tag{1.1}$$

However, as each observation is not equally likely to be selected, the terms in this sum relating to each observation in Equation 1.1 are instead inversely weighted by their probabilities of having been sampled, π_i. Thus we solve

$$\sum_{i=1}^{N} \frac{R_i}{\pi_i} \frac{\partial}{\partial \beta} \log(f(y_i|\beta, \mathbf{x}_i)) = 0,$$

where R_i is an indicator for whether or not possible observation i was included in the sample, and i now ranges over all possible observations. R_i may be thought of as following a Bernoulli distribution with probability π_i. Such a method is used in survey sampling, the principle behind it being that if we have an observation which is from an under-represented group (i.e. with a low value of π_i) we will give more weight to that value in the log-likelihood, in order to adjust for the under-representation. A similar method proposed is the 'conditional maximum likelihood' method, which involves solving

$$\sum_{i=1}^{N} R_i \frac{\partial}{\partial \beta} \log\left(f(y_i|\beta, \mathbf{x}_i, R_i = 1)\right) = \sum_{i=1}^{N} R_i \frac{\partial}{\partial \beta} \log\left(\frac{f(y_i|\mathbf{x}_i, \beta)\pi(y_i, \mathbf{x}_i)}{\int f(y_i|\mathbf{x}_i, \beta)\pi(y_i, \mathbf{x}_i) dy_i}\right) = 0,$$

i.e. maximising the conditional log-likelihood, given that the pair $(y_i, \mathbf{x}_i)$ was observed, which is essentially another method of weighting each observation according to the probability of their being in the sample. Both such methods require us to have knowledge of, or the ability to estimate these inclusion probabilities. These methods, while dealing with an analogous situation in which the aim is to adjust for under-sampling in certain groups (rather than spatial regions) are not immediately applicable to our situation, as we do not know what these probabilities of inclusion are.

Zidek, Shaddick and Taylor (2014)

Zidek et al. (2014), inspired by such data as provided by the UK Black Smoke network, extend the weighting methods of Scott and Wild (2011) to the field of air pollution networks by using the Horwitz-Thompson estimator. They consider a finite set of locations u_j with $j = 1, \ldots, N$ at which samples Y_{jt} may be taken at times $t = 1, \ldots, T$. They also include covariates $\mathbf{X}_j$, which may include things such as the classification of a site into an urban or rural area, or how close a site is to a major road. It is assumed that the selection process $\{R_i : i \in 1 \ldots N\}$, is stochastic with uncertain probabilities, which may be dependent on both Y and $\mathbf{X}$ at previous times. The probability of the inclusion of any site at time t is thus modelled on Y and $\mathbf{X}$ observed at times $1, \ldots, t-1$. They demonstrate that when such a time series of inclusion (or not) of monitors is available, this is an effective method for adjusting for response-dependent monitor closures. The main focus here is on official statistics, such as mean pollutant concentrations, rather than mapping.

Watson, Shaddick and Zidek (2019)

Watson et al. (2019) propose a spatio-temporal model in which latent effects specific to time and space are included in both the linear predictor for the observation process and the predictor of any particular site's operationality at a particular time. These shared latent effects may include any combination of i.i.d. random effects, specific to one site or time, or spatially or temporally correlated random effects. Meanwhile, both processes also contain fixed effects based on the (possibly different) fixed covariates. In their modelling of Black Smoke data this probability of a site being operational depends on the value of these effects at previous times, rather than the value of the underlying process itself at the current time. This very general framework brings together many of the existing models: when purely spatial data (rather than a time series of spatial data) is available, shared covariates predict both the surface and the probability of any single site's operationality. However, when temporal data is available, as for the black smoke data, operationality may be modelled from previous times. Additionally, as with other methods, it is assumed that, given underlying processes upon which they are based, the selection probability of any particular site is independent of the selection probabilities of the other sites.

1.3 Utility functions and experimental design

Our approach to dealing with preferential sampling will be based on the notion of modelling the preferences of an experimenter, jointly with our modelling of the process of interest, via uitlity functions to describe the utility of any particular design to the experimenter. Utility functions are a tool used in decision theory and risk analysis to quantify systematically the preferences of an individual by giving a ranking to a set of possible elements to be chosen from. More on these in a general context can be found in Lindley (1975) and Gerber and Pafum (1998), among others. In the context of experimental design, an experimenter is likely to have objectives in terms of the outcome of the experiment which will influence the experimental design choices. Naturally, these do not need to be made by consciously maximising a function: for example, designs which have an even spread of monitors have beneficial properties in terms of minimising prediction variance, as discussed in Nychka et al. (1997), as is intuitive if one wants to get a good picture of a whole region.

Müller (2005) describes how experiments (not necessarily geospatial) may be designed optimally by maximising an explicit utility function by which the experimental preferences are encapsulated. Given the observable process y, parameters θ, possible design d, conditional distribution of the observable process $p_d(y|\theta)$, and prior distribution $p(\theta)$, such optimal designs d^* may be estimated by maximising the expected value of the utility function:

$$d^* = \arg\max_d \int U(d, y, \theta) p_d(y|\theta) p(\theta) d\theta dy.$$

Clearly this requires at least some prior knowledge about the parameters θ and how they are related to the observable process y. They describe how, given this prior knowledge, simulation based techniques may be used to estimate d^*. They consider utility functions such as those which minimise the bias in parameter estimates prediction error, among others.

da Silva Ferreira and Gamerman (2015) bring this concept of utility based designs into the realm of geostatistics by describing methods for determining optimal placing of new sampling sites which are to be added to existing sampling networks. They describe how finding such an optimal placement involves maximising or minimising a function which 'quantifies the gains and losses related to each possible decision', i.e. a utility function. (Interestingly, they show that this decision process is substantially affected by any preferential sampling already present in the network). They consider functions which minimise the expected prediction variance in the final predictions, given the current values output by the sampling network. Additionally, they propose utilities which are higher for locations s in which extreme values of the observable process Z are to be expected, given current observed values y and parameters θ. For example the utility

$$U(s_d, \theta, y) = P(Z(s_d) > a|\theta, y), \tag{1.2}$$

is proportional to the probability that the value $Z(s_d)$ measured at chosen location s_d is higher than some threshold value a.

The methods described in this section deal predominantly with the addition of new, individual sampling sites to existing networks which are already producing observations. In that which follows we shall present methods by which sampling designs may be modelled as a whole via utility functions, allowing for interactions between sampling sites, in order to better adjust for preferential sampling in the inference on the measured process of interest. Additionally, while these utility based methods deal predominantly with the task of designing a sampling network, we shall instead employ them post-hoc, in order to perform analysis of the underlying field.

1.4 Our approach

There are two main avenues of enquiry when adressing preferential sampling, that of correcting for bias in parameter estimation, and that of predicting the surface Z. We shall focus on the latter, in a spatial (rather than spatio-temporal) setting.

The models of Diggle et al. (2010), Pati et al. (2011) and Gelfand et al. (2012), described above, all treat the sampling process as an inhomogeneous Poisson process of some kind. Implicit within these processes is the assumption that, given the intensity, the sampling locations are independent of one another. There are indeed situations in which this may be a valid assumption, for example, in situations in which there has not been a prespecified design, such as the University of Georgia's Marine Debris Tracker (Jambeck and Johnsen (2015)) which allows members of the public to report locations of found marine debris. Once factors such as the desirability of travelling to a particular marine location, popularity of the tracker, and marine pollution density (which may or may not be independent of one another) have been accounted for it is unlikely that a sighting of debris in one location should be dependent on the sighting at another.

However, when pre-planned sampling networks are designed as a whole, this becomes an unrealistic assumption. Say we had two large regions of roughly equal interest. Once we have decided to place a monitor in one of them, it may become more desirable to place the next in the other, than to put

another in the same region, in order to achieve broader coverage. Likewise, aside from any value related preference, an experimenter may favour a gridded, or space-filling sampling design. We shall investigate the possibility of using a 'whole-design utility' to encapsulate the usefulness of whole designs to the experimenter. This enables us to include dependence between sampling locations, allowing for realistic modelling for a wider range of preference, such as space-filling of a particular region, or repulsion effects between monitors etc.. We shall use the following definitions:

Sampling Design: A sampling design on a spatial domain S of size n is a random discrete subset of size n of S. We denote designs by $\tilde{D}$ and the space of designs on S by $\mathcal{D}_n$.

Gaussian Spatial Process: A Gaussian spatial process $\tilde{Z}$ on a spatial domain S is a random function on S with finite dimensional Gaussian distributions. We denote the space of all Gaussian spatial processes on S by $\mathcal{Z}$.

Design Utility: A design utility $\tilde{U}$ is a function $\tilde{U} : \mathcal{D}_n \times \mathcal{Z} \to (0, \infty)$.

As the spatial domain S is continuous, we discretise it in order to fit the Gaussian Process in question. Let $C = \{C_1, \dots C_N\}$ be a partition of the spatial domain S and $\mathbf{c} = (c_1, \dots, c_N)$ with $c_i \in C_i$ a 'focus point' of C_i. We model the utility of a design $\tilde{D}$ and a Gaussian Process $\tilde{Z}$ based on C, $\mathbf{c}$ and parameters α as

$$\tilde{U}(\tilde{D}, \tilde{Z}; C, \mathbf{c}, \alpha) = U(D, Z; \alpha),$$

where D is a length N vector of counts $D = (d_1, \dots, d_N)$, where $d_i = |\tilde{D} \cap C_i|$ is the number of elements in $\tilde{D} \cap C_i$, and Z is a vector of length N such that the i^{th} element $z_i = \tilde{Z}(c_i)$. We denote the space of all count vectors of a design of size n and a partition of size N by $\mathbb{D}_{N,n}$. This space contains $\binom{N+n-1}{n}$ distinct elements $D = (d_1, \dots, d_N)$ with $d_i \in \mathbb{N} \quad \forall i \in \{1, \dots, N\}$ and $\sum_{i=1}^{N} d_i = n$ [1]. This setup is illustrated in Figure **??**. Utility functions will, in general, be described in terms of these elements d_i $i \in \{1, \dots, N\}$ of D. Partitioning the space is, in practice, necessary for the prediction of a Gaussian spatial process as we must define a set of points at which to predict its value. It makes practical sense (e.g. when adjusting these predictions for preference at these prediction locations) to associate these locations with the number of sampling points in a small neighbourhood or cell around them, as defined by the partition. Consequences of the choice of partition will be discussed in Chapter 6. The function U is a utility function such that

$$U : \mathbb{D}_{N,n} \times \mathbb{R}^N \to (0, \infty).$$

Generally we will drop the N and n subscripts. Such utility functions may also depend on one or more parameters α etc., which may be used to specify the strengths of preferences held by an experimenter.

[1] This is the value of the multiset coefficient for n choices from N items when items may be chosen multiple times. This may be demonstrated by considering that there is a bijection between the design space and the space of n elements separated by $N-1$ 'separators', between which are the number of elements corresponding to each of the N cells (no separator is necessary before any elements corresponding to the first cell, or after any elements corresponding to the last cell). The number of these element-separator arrangements is simply the binomial coefficient $\binom{N+n-1}{N-1} = \binom{N+n-1}{n}$, as there are now $N+n-1$ spaces in which the $N-1$ separators can be placed.

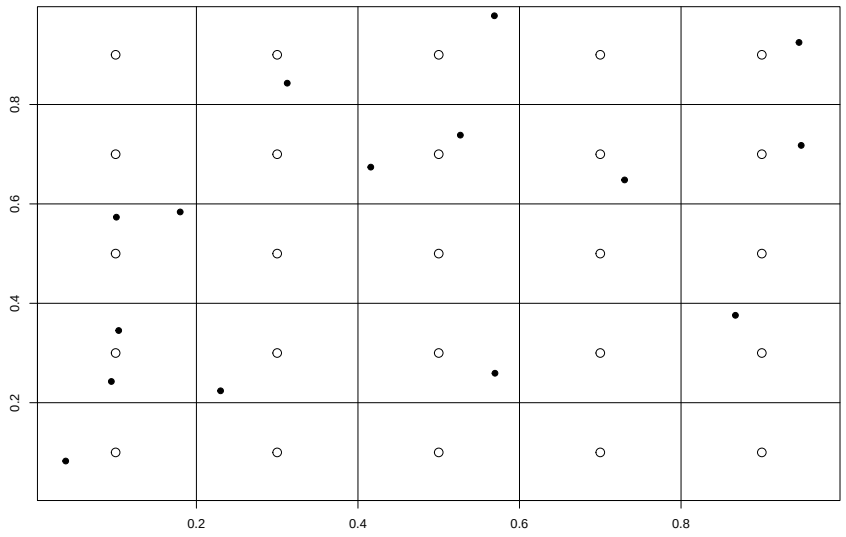


Figure 1.2: In the above, the large square represents the spatial domain S. The grid cells are the partitions $C_1, \ldots C_N$ with the unfilled dots corresponding to the focus points $c_1 \ldots c_N$. The sampling design $\tilde{D}$ corresponds to the filled dots. The count vector $D = (0, 1, 1, 0, 1, 0, 0, 2, 1, 1, 2, 0, 0, 0, 0, 2, 1, 1, 0, 1, 1, 0, 0, 0, 0)$ starting from the top left cell and moving horizontally first and downwards. This in turn would correspond to values $d_1 = 0$, $d_2 = 1$, $d_3 = 1$, $d_4 = 0$ etc.. The vector Z would correspond to the values of the Gaussian random field at the focus points (unfilled dots), while the vector X would correspond to the values of the Gaussian random field at the filled dots. The measurements Y are assumed to be these values X with added mean-zero measurement error.

These preferences may be for the detection of higher values of the process of interest, or for a good coverage of the spatial region etc..

We will assume that the probability of selection of any D depends only on the values of $\tilde{Z}$ at $\mathbf{c}, Z$, and is proportional to its utility:

$$P(D|Z;\alpha) \ \propto \ U(D, Z;\alpha).$$

This model assumes that if D is fixed then all designs $\tilde{D}$ such that $|\tilde{D} \cap C_i| = d_i$ are equally likely, i.e.

$$P(\tilde{D}|D) \propto \prod_{i=1}^{N} 1(|\tilde{D} \cap C_i| = d_i).$$

Therefore, given the number of samples from each partition element, we assume that the monitors are then uniformly distributed within those elements of the partition. This assumption is based on the notion that this value $z_i = \tilde{Z}(c_i)$, the value of the continuous process $\tilde{Z}$ at the focus point c_i of cell i, is reasonably representative of the of the values of the process within that cell. When using arbitrarily defined squares of size larger than the scale of the variation of Z, this may not be a realistic assumption. The effects of this assumption are discussed further in Chapter 6, in which the Z-related arguments of the utility functions may be the mean values of $\tilde{Z}$ at the focus points of many cells, assumed to be

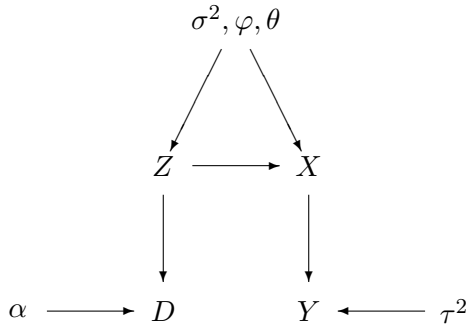


Figure 1.3: DAG displaying the dependence structure in our proposed model.

representative of the mean value of $\tilde{Z}$ over those cells.

Our approach of using a whole-design utility provides a more flexible framework for encapsulating the intentions of the experimenter and allows for both dependent sampling locations, and mixed preferences such as a balance between a preference for higher values, and an optimal coverage of the whole region. Given a very fine discretisation of the region, the use of a utility function such that the probabilities of selection of each site are independent and proportional to the exponential of a multiple of the value of the process within them, is analogous to the method of Diggle et al. (2010).

In practice, it is likely that an experimenter would select the highest-utility design with certainty, rather than by drawing from a distribution. However, such a framework allows us to encapsulate our uncertainty with respect to the sampling process, as we do not have access to perfect information about the sampling process, which in practice will be based on a vast array of not easily quantified reasoning, and because there will be much uncertainty about the surface $\tilde{Z}$ before the monitors are actually placed.

1.5 Implementation details

We aim to build a framework for predicting a surface, given preferentially-sampled spatial measurements of that surface, where there is an unknown level of preference for high or low values. As described above, our approach will involve discretising the spatial region by disjoint rectangles, called cells, giving a finite number of possible designs, selected according to a probability distribution proportional to its utility. The strength of experimenter preferences will be controlled by a parameter, or parameters α, to be estimated from the data. Our model will assume that the selection of any grid cell for a site is dependent on the underlying surface only via its value at the centre of the cell, the focus point, and the positioning of the sampling site within the square will follow a normal distribution. We will thus estimate $\tilde{Z}$ at two sets of locations: Z, of length N at the centres of the grid squares, the focus points, and X, of length n, of $\tilde{Z}$ at the actual sampling sites, i.e. $Z = \tilde{Z}(\mathbf{c})$ and $X = \tilde{Z}(\tilde{D})$. A network showing the dependency structures is displayed in Figure 1.3, in which φ, σ^2 θ are parameters for the Gaussian random field, τ^2 is the measurement noise parameter, Y the measurements, and α the strength of preference parameter.

The joint likelihood of Z, X, D and Y may be factorised, given parameters, as

$$P(Z, X, D, Y) = P(Y|X)P(X|Z)P(D|Z)P(Z), \tag{1.3}$$

where

$$P(D|Z) = \frac{U(D, Z; \alpha)}{K(Z, \alpha)}, \tag{1.4}$$

with

$$K(Z; \alpha) = \sum_{D \in \mathbb{D}} U(D, Z; \alpha),$$

which requires a summation over over all $\binom{N+n-1}{n}$ possible designs, and is thus often incalculable. This means that we are unable to evaluate the likelihood in question, or, in a Bayesian context, the posterior distribution of Z. One of the key challenges which we will address will be the calculation of a suitable approximation to this normalising constant, or rather, the ratio of two such normalising constants. For example, we would require $\frac{K(Z,\alpha)}{K(Z^*,\alpha^*)}$ in a maximum-likelihood context in which Z^* and α^* are some fixed reference values, or $\frac{K(Z_1,\alpha_1)}{K(Z_2,\alpha_2)}$ where Z_1 and α_1, and Z_2 and α_2 are values to be chosen between, with probabilities determined by a Metropolis-Hastings ratio in a Markov chain Monte Carlo context. This, would involve a proposal Z^* at step t of the algorithm, with current value Z^t, from proposal distribution $Q(Z|Z^t)$. This new value Z^* must then be accepted with probability

$$\min\left(1, \frac{P(Z^*)P(X|Z^*)U(D, Z^*; \alpha)K(Z^t, \alpha)Q(Z^t|Z^*)}{P(Z^t)P(X|Z^t)U(D, Z^t; \alpha)K(Z^*, \alpha)Q(Z^*|Z^t)}\right),$$

which contains the ratio of incalculable normalising constants. This MCMC scenario will be our primary consideration.

1.6 Contributions of this thesis

The main new contributions of this research are the use of highly generalisable utility functions which can be used to model an arbitrarily wide range of experimental preferences, allowing for more realistic preference estimation, and better inferences of the process of interest. To our knowledge, this is the first instance in which sampling designs are modelled as a whole, allowing for interactions between sampling locations, in order that preferential sampling may be accounted for. We demonstrate how taking the design as a whole, rather than as a collection of independent sampling locations can lead to improved inferences on the process of interest. This is achieved by specifying utilities for the designs that encapsulate the preferences of the experimenter.

We begin by reviewing the use of a utility function which corresponds to a multinomial distribution for the elements of the design (Chapter 2), to demonstrate the effects of preferential sampling and the importance of accounting for it. This is analogous to the work of Diggle et al. (2010). Following this, we broaden the scope of preferences accounted for, and consider a range of possible utility functions, discussing how we might choose one which appropriately captures the intentions of an experimenter. In particular, we shall focus on the situation in which a sampling network exhibits a balance between good coverage of the region of interest, and a preference for higher values.

The selection of a utility function for modelling good coverage is a highly nontrivial problem: we shall demonstrate that pairwise distance functions, and the minimax and maximin criteria, which are

commonly used to find space-filling designs, are not useful when there is a balancing of preferences, and present more useful alternatives. This balanced preference for good coverage and high values represents an important challenge: these preferences can be thought of as working in 'opposite directions': one favouring clustering and the other penalising clustering. If the coverage preference is ignored, then a level of clustering in a high valued region, which would not be out of the ordinary in a uniform sampling scenario (while very much out of the ordinary in an optimal coverage design scenario), will go undetected. Consequently, the preferential sampling will be underestimated, leading to inaccurate inferences. We demonstrate that utilities based on average distances rather than minimum and maximum distances are effective in encapsulating coverage related preferences: this is explored in Chapter 3.

As discussed, a major challenge is presented by the ratio of intractable normalising constants associated with the design distributions implied by the utility functions in question. We will compare different methods of approximating the ratio of normalising constants, some of which require design samples drawn from $P(D|Z, \alpha)$. We will also present methods for generating and evaluating such samples, in addition to presenting a class of utility functions, defined as 'permutation invariant utility functions' which exhibit interesting mathematical properties that facilitate more efficient estimation of the ratio of normalising constants (Chapter 4). We will show that the estimation of the normalising constants relating to these particular utility functions often requires substantially fewer samples to reach a similar accuracy to those without the relevant properties. We incorporate these techniques into a wider Bayesian inferential framework in which the generation of samples of designs is used for the estimation of normalising constants for calculating the acceptance probability of a Metropolis Hastings algorithm. To improve the estimation of the high-dimensional vectors Z, we will describe and make use of an approximate, or 'noisy' Metropolis adjusted Langevin algorithm, which will also require the estimation of these normalising constants and the generation of samples.

The methods developed will be applied to several data sets, both simulated and real, such as a set of soil Ammonia concentrations taken from a field in Scotland (Chapter 5), or measurements of Nitrogen Dioxide air pollution in the city of Bath (Chapter 2). We shall consider the effects of the choice of discretisation on the design distributions and the predicted surfaces in Chapter 6.

Finally, we shall consider how multiple sources of data might be employed to give better predictions of preference levels, demonstrating such strategies on data sets of Lead pollution levels from moss samples in Galicia, Spain (Chapter 7). Such methods of combining data sets are promising, in that they may be a valuable tool for integrating data from sources with different sampling schemes: e.g satellite and on-the-ground pollution measurement, or disease prevalence mapping where there is a large number of asymptomatic cases, and the disease may be detected from both symptomatic cases and environmental surveillance. We show how the combination of data from different sources, and the resulting improved preference estimation can make both data sets more informative than if they had been used in isolation.

Chapter 2

Multinomial utility functions

We begin by considering the use of a utility function which corresponds to a multinomial distribution for the design, that is, having partitioned the space into cells, we assume that the probability that any monitor is placed within a particular cell is independent of where the other monitors are placed, and proportional to $\exp(\alpha z_i)$ where $z_i = \tilde{Z}(c_i)$ is the value of the process of interest associated with cell i, and α an (unknown) parameter which encapsulates the strength of preference of the experimenter. This is analogous to the log-Gaussian Cox process method of Diggle et al. (2010). This whole-design utility of a design $D = (d_1, \ldots, d_N)$, for n monitors placed in N cells, and resulting in cell counts d_i, $i \in 1, \ldots, N$, and Gaussian random field values $Z = (z_1, \ldots, z_N)$, may be expressed as

$$U(D; Z, \alpha) = \frac{n!}{\prod_{i=1}^{N} d_i!} \frac{\exp(\alpha \sum_{i=1}^{N} z_i d_i)}{(\sum_{j=1}^{N} \exp(\alpha z_j))^n} = \frac{n!}{\prod_{i=1}^{N} d_i!} \prod_{i=1}^{N} \left(\frac{\exp(\alpha z_i)}{\sum_{j=1}^{N} \exp(\alpha z_j)} \right)^{d_i}. \tag{2.1}$$

The normalising constant of this utility function, as it is simply a multinomial distribution, is equal to 1. Defining a_i to be the area of cell i, we have, in terms of the continuous space design $\tilde{D}$,

$$P(\tilde{D}|Z, \alpha) = \sum_{D} P(\tilde{D}|D) P(D|Z, \alpha),$$

where

$$P(\tilde{D}|D) = \begin{cases} \frac{\prod_{i=1}^{N} d_i!}{\prod_{j=1}^{N} a_j^{d_j}}, & \text{if } \tilde{D} \in D, \\ 0 & \text{otherwise,} \end{cases} \tag{2.2}$$

where we say $\tilde{D} \in D$ if and only if $d_i = |\tilde{D} \cap C_i|$ for all $i = 1, \ldots, N$. The denominator of this follows from the fact that the area within which each monitor can be placed, given that it is within cell j, is a_j. The numerator follows from the fact that, given d_i locations at which monitors are placed within cell i, there are $d_i!$ ways in which the monitors can be arranged, indistinguishably, within them. Putting these components together gives

$$p(\tilde{D}|Z, \alpha) = \frac{n!}{\prod_{i=1}^{N} a_i^{d_i}} \frac{\exp(\alpha \sum_{i=1}^{N} z_i d_i)}{(\sum_{j=1}^{N} \exp(\alpha z_j))^n}.$$

.

Parameter	True value	Preferential mean (s.d.)	Non-preferential mean (s.d.)
α	2.15	1.98 (0.325)	NA
τ^2	0.01	0.0487 (0.0680)	0.0468 (0.0650)
σ^2	3	2.43 (2.23)	1.93 (1.05)
$\log(\varphi)$	-0.15	0.0217 (0.713)	-0.251 (0.568)
θ	50.0	50.2 (2.10)	52.4 (4.77)
Sum of squared errors	-	196	446
Mean absolute errors	-	0.338	0.696
Mean squared errors (low)	-	0.807	2.047
Mean squared errors (high)	-	0.175	0.185
Mean absolute errors (low)	-	0.733	1.28
Mean absolute errors (high)	-	-0.0563	0.109
Proportion overestimated	-	279/400	319/400

Table 2.1: Parameter estimates and sum of squared errors for simulated preferentially-sampled data and preferential and non-preferential model fits. The mean squared and absolute errors are reported for the locations corresponding to the values of Z above and below its median: (high) and (low) respectively.

2.1 Exploring the effects of preferential sampling

We carry out several experiments to demonstrate both the effects of preferential sampling and the mitigating effects of including the sampling process in the prediction model. To this end, we generate an example data set using a simulated Gaussian random field and its realisation Z over a regular 20×20 grid on the unit square. We choose this Gaussian random field to have mean $\theta = 50$, variance $\sigma^2 = 3$, and exponential correlation function $C(s_i, s_j) = \exp\left(\frac{-d_{ij}}{\varphi}\right)$ for locations s_i and s_j at a distance of d_{ij} from one another, with correlation parameter φ, where $\log(\varphi) = -0.15$. Sampling cells are chosen via the multinomial utility function (2.1), with strength of preference parameter $\alpha = 2.15$. We then select $n = 20$ irregular sampling locations uniformly within the chosen cells, and add Gaussian noise with variance $\tau^2 = 0.01$ to the measured values X of the Gaussian random field at those locations, to give measurements Y. In order to demonstrate the advantages of accounting for the sampling process, we now attempt to reconstruct Z from these measurements. We fit two models, building into the first the assumption that preferential sampling has taken place, while in the second assuming that there has been no preferential sampling. The relationships in this model are the same as those displayed in the graph in Figure 1.3. We set inverse gamma priors for φ, σ^2 and τ^2 with hyperparameters $(2, 1)$, $(2, 4)$ $(2, 0.05)$ respectively. We assign Gaussian priors to θ and α with prior means and variances $(52, 10)$ and $(1, 50)$ respectively. Secondly, using the same parameters and original Gaussian random field, we repeat the same experiment, for comparison, with uniformly distributed sampling locations, in order demonstrate the effects of preferential sampling. We fit the model via MCMC with 100000 iterations, the first 25000 of which are discarded as burn-in. Comprehensive details of the fitting algorithm, and comparisons with other models are given in Chapter 5.

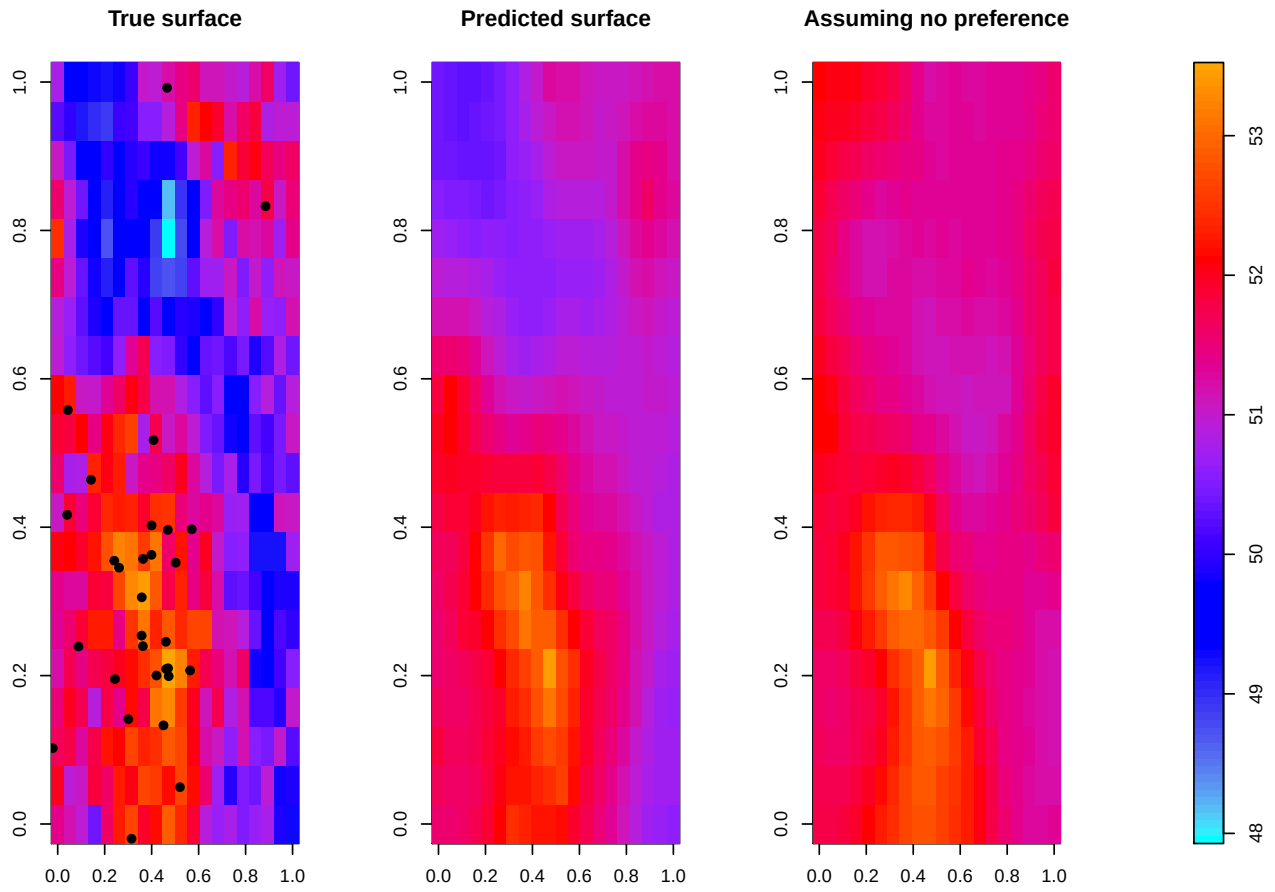


Figure 2.1: Original and predicted surfaces, where we assume preferential sampling (centre) and assume uniform sampling (right). The sampling locations are superimposed on the original surface (left).

Parameter	True value	Preferential mean (s.d.)	Non-preferential mean (s.d.)
α	-	0.00919 (0.234)	NA
τ^2	0.01	0.0483 (0.0663)	0.0468 (0.650)
σ^2	2	2.47 (1.83)	2.41 (1.38)
$\log(\varphi)$	-0.15	-0.484 (0.633)	-0.450 (0.526)
θ	50.0	50.6 (1.50)	51.5 (1.42)
Sum of squared errors	-	125	117
Mean absolute errors	-	0.0767	0.0993
Mean squared errors (low)	-	0.348	0.321
Mean squared errors (high)	-	0.276	0.266
Mean absolute errors (low)	-	0.333	0.314
Mean absolute errors (high)	-	-0.180	-0.115
Proportion overestimated	-	222/400	230/400

Table 2.2: Parameter estimates and sum of squared errors for simulated uniformly sampled data and preferential and non-preferential model fits. The mean squared and absolute errors are reported for the locations corresponding to the values of Z above and below its median: (high) and (low) respectively.

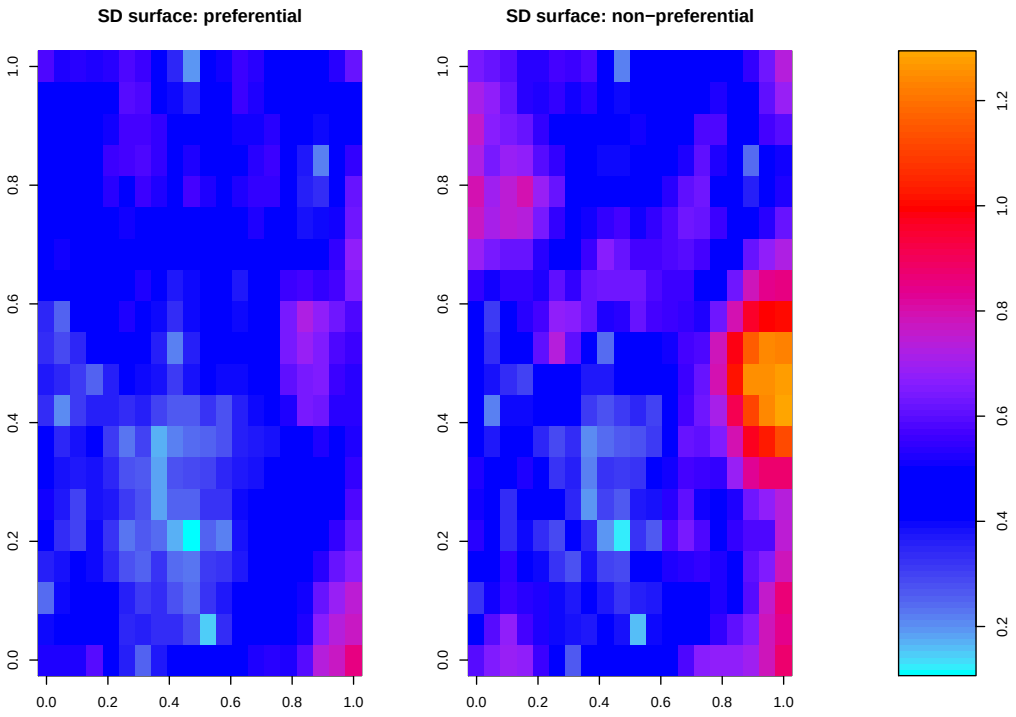


Figure 2.2: Standard deviation surfaces for surfaces, where we assume preferential sampling (left) and assume uniform sampling (right).

Results

These four experiments (two simulated data sets: preferential and non-preferential, two models: preferential and non preferential) demonstrate several ideas. We compare the posterior mean estimates for Z by considering the sum of squared errors, which gives an overall picture of how far away from the true values of Z the predicted values of Z are, penalising bigger errors with a larger weighting. Additionally, to demonstrate which elements (i.e. higher or lower) of Z are underestimated or overestimated, we consider the mean absolute errors (positive values indicating overestimation) for all elements of Z, those at locations which correspond to the values of Z below its median, and those at locations which correspond the values above the median. We also consider the proportion of values which have been overestimated. These metrics of comparison and estimated parameter values can be seen in Tables 2.1 and 2.2. Trace and density plots are shown in Figures 2.3, 2.4, 2.5, 2.6 for the preferential data case, and 2.10, 2.11, 2.12, 2.13 for the non-preferential data. The first thing we note, for this situation, is that is Z is predicted best where there has not been preferential sampling, regardless of the fitting model: in terms of the sum of squared errors (125 and 117 opposed to 196 and 446), and in terms of the mean absolute errors, showing that there is much less overestimation, especially for the lower values of Z. This is to be expected, as the lower values are much less sampled in the preferential sampling situation. Similarly, fewer values are overestimated where the data is sampled uniformly. While in both situations the preferentially sampled data leads to worse estimation than if the data has been sampled uniformly, we can see that the situation is much worse when we do not take the preferential sampling into account:

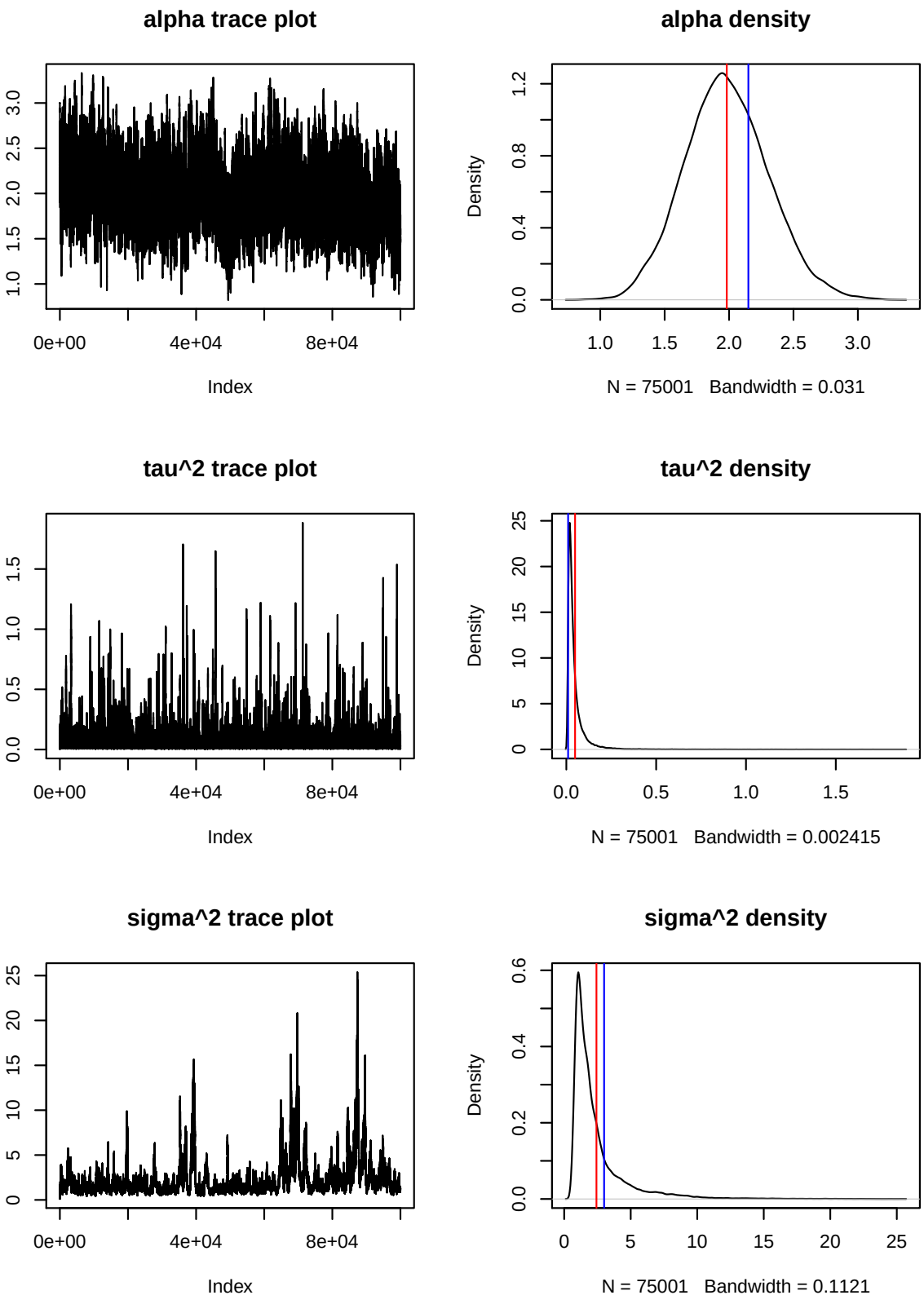


Figure 2.3: Parameter plots for the preference-assumed model fit to the simulated preferentially-sampled data. Red lines show predicted values, blue lines show actual values.

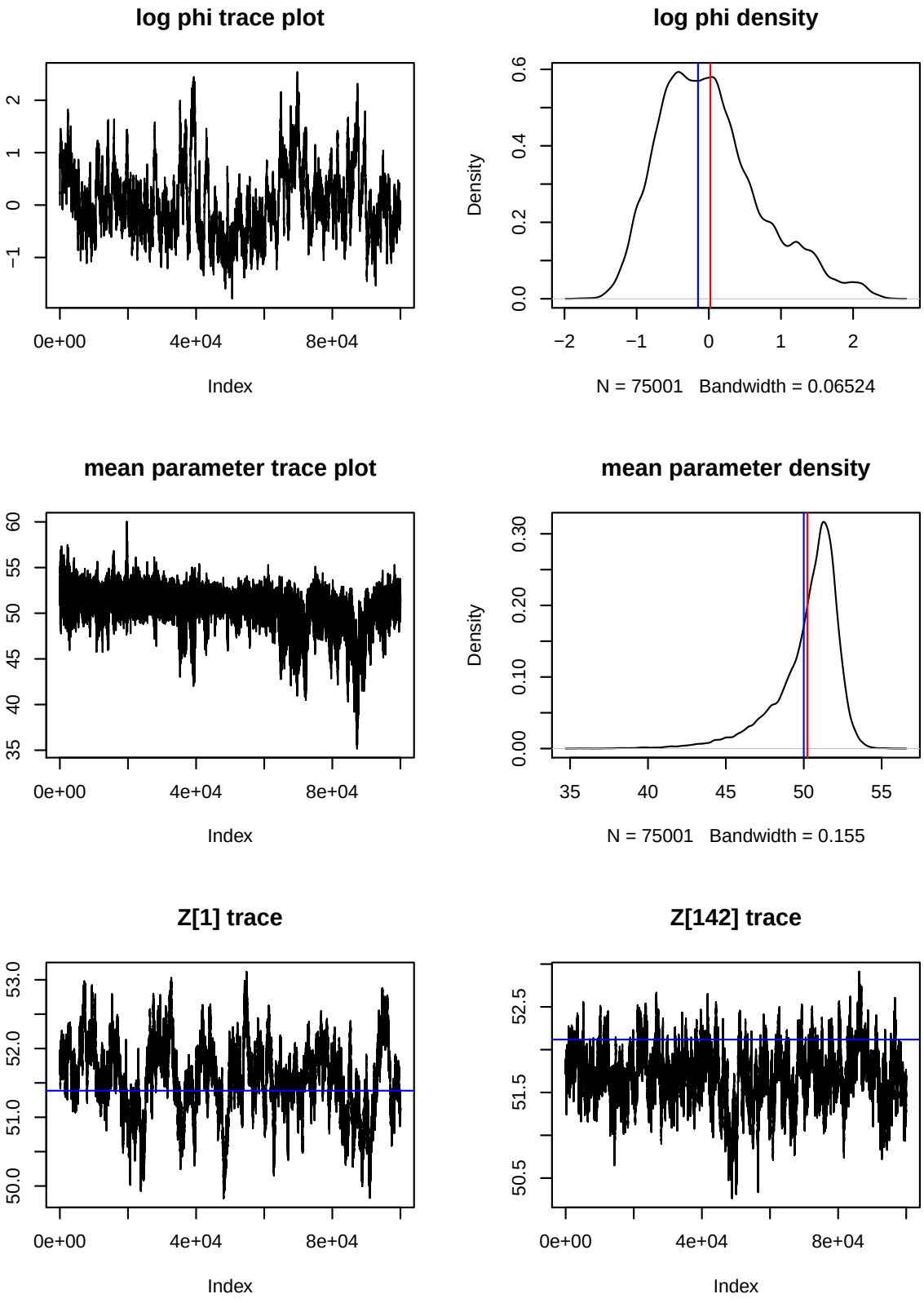


Figure 2.4: More parameter plots for the preference-assumed model fit to the simulated preferentially-sampled data, along with two trace plots for two selected Z values. Red lines show predicted values, blue lines show actual values.

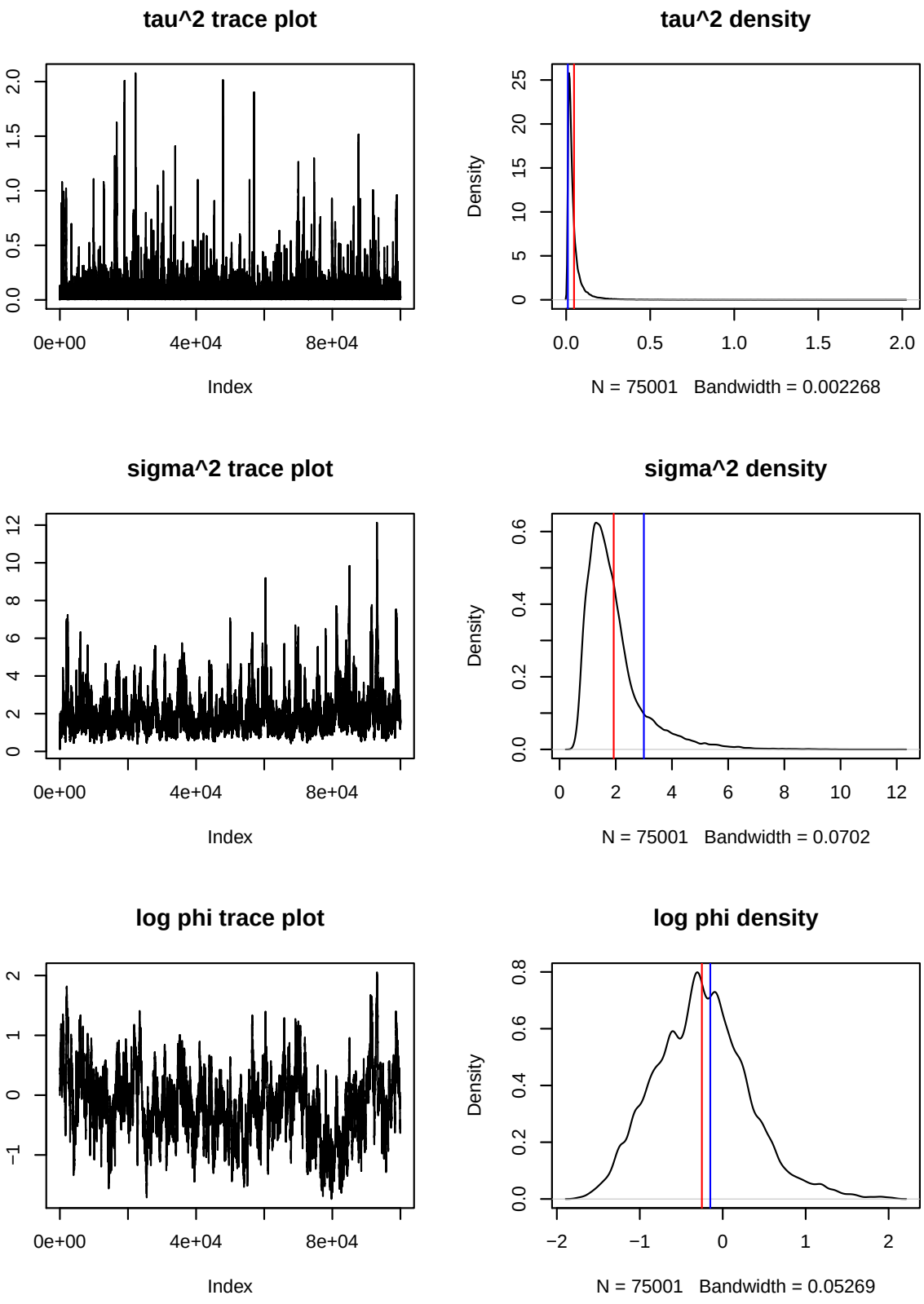


Figure 2.5: Parameter plots the non-preferential model fit to the simulated preferentially-sampled data. Red lines show predicted values, blue lines show actual values.

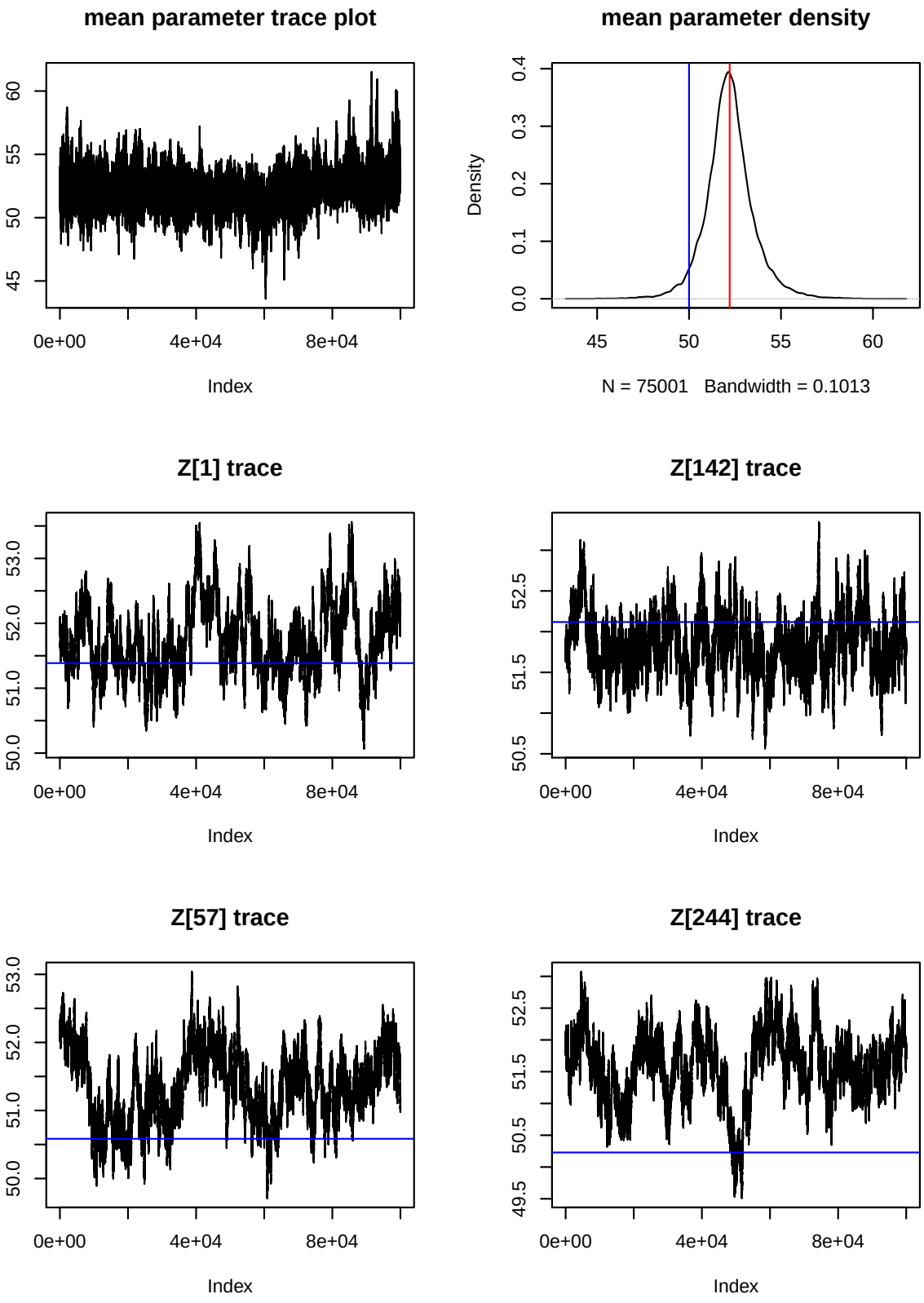


Figure 2.6: More parameter plots the non-preferential model fit to the simulated preferentially-sampled data, along with trace plots for selected Z values. Red lines show predicted values, blue lines show actual values.

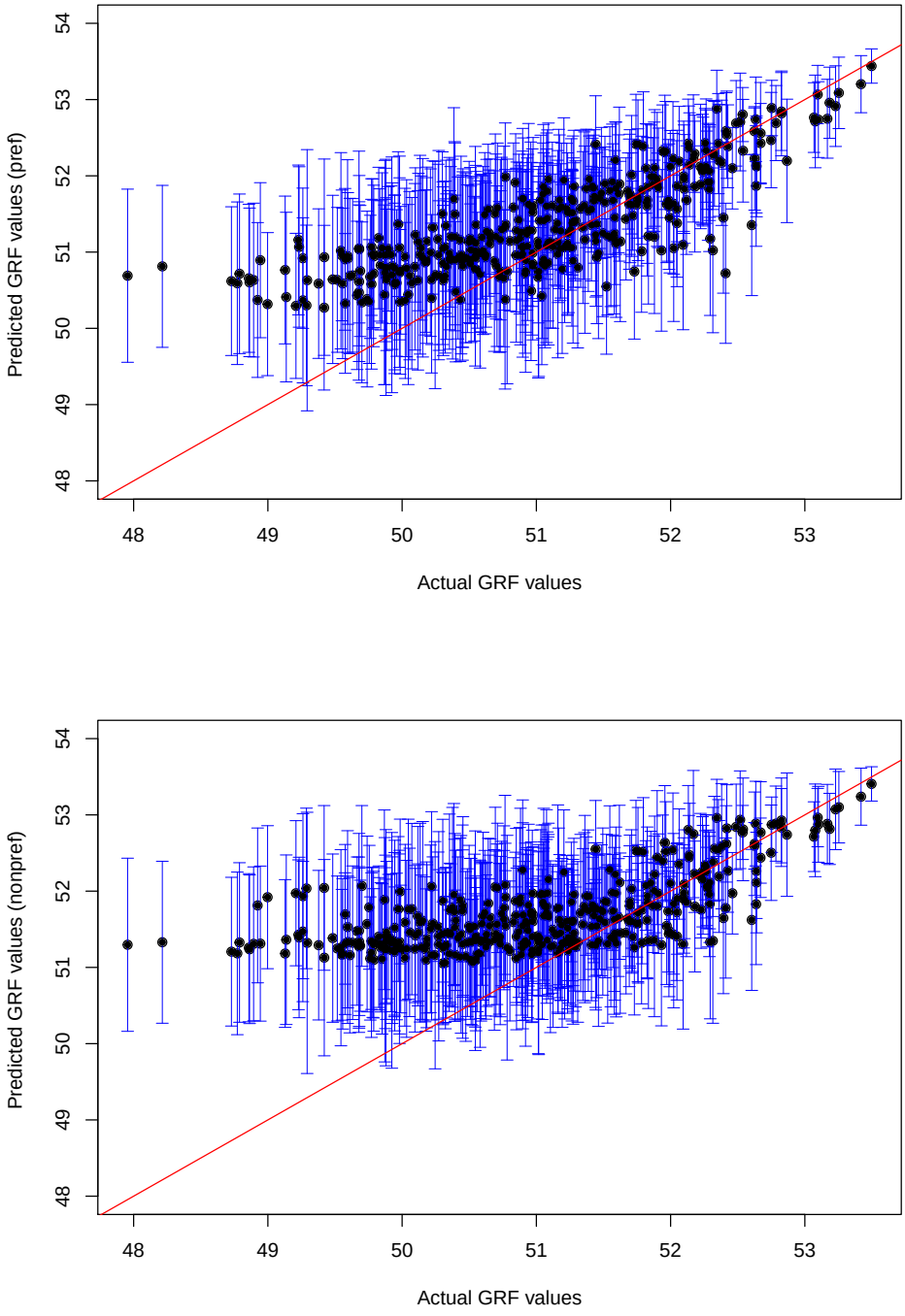


Figure 2.7: Predicted values against actual values for the simulated preferentially-sampled data, for both methods of prediction, both assuming and ignoring preference. Error bars are one standard deviation either side.

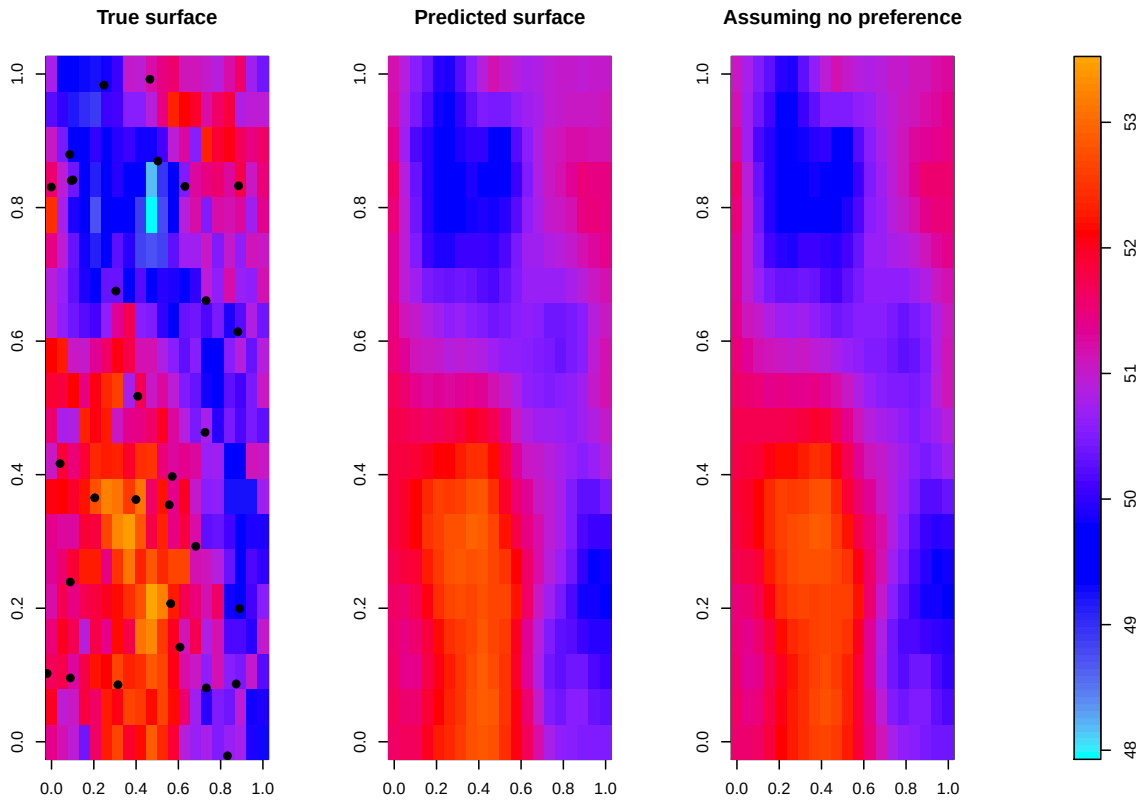
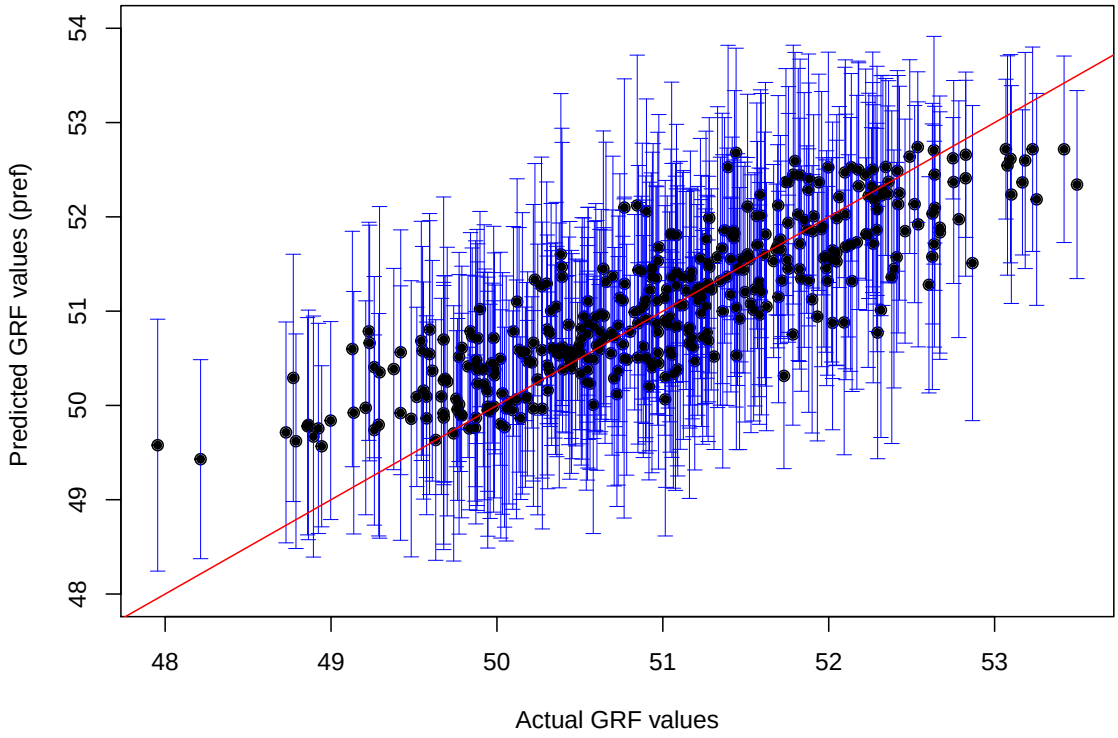


Figure 2.8: Actual and predicted surface, for non-preferentially-sampled data, but a model which allows for preference, and one which does not .



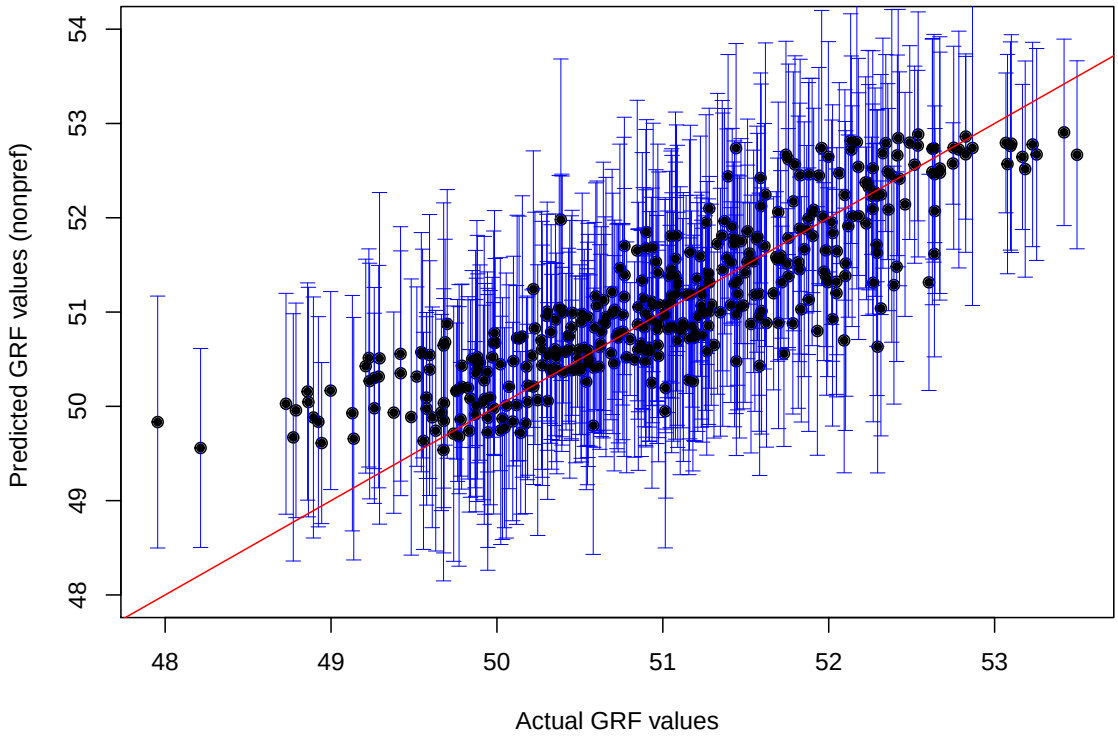


Figure 2.9: Actual and predicted Z values or non-preferentially sampled data, but a model which allows for preference. Error bars are one standard deviation either side.

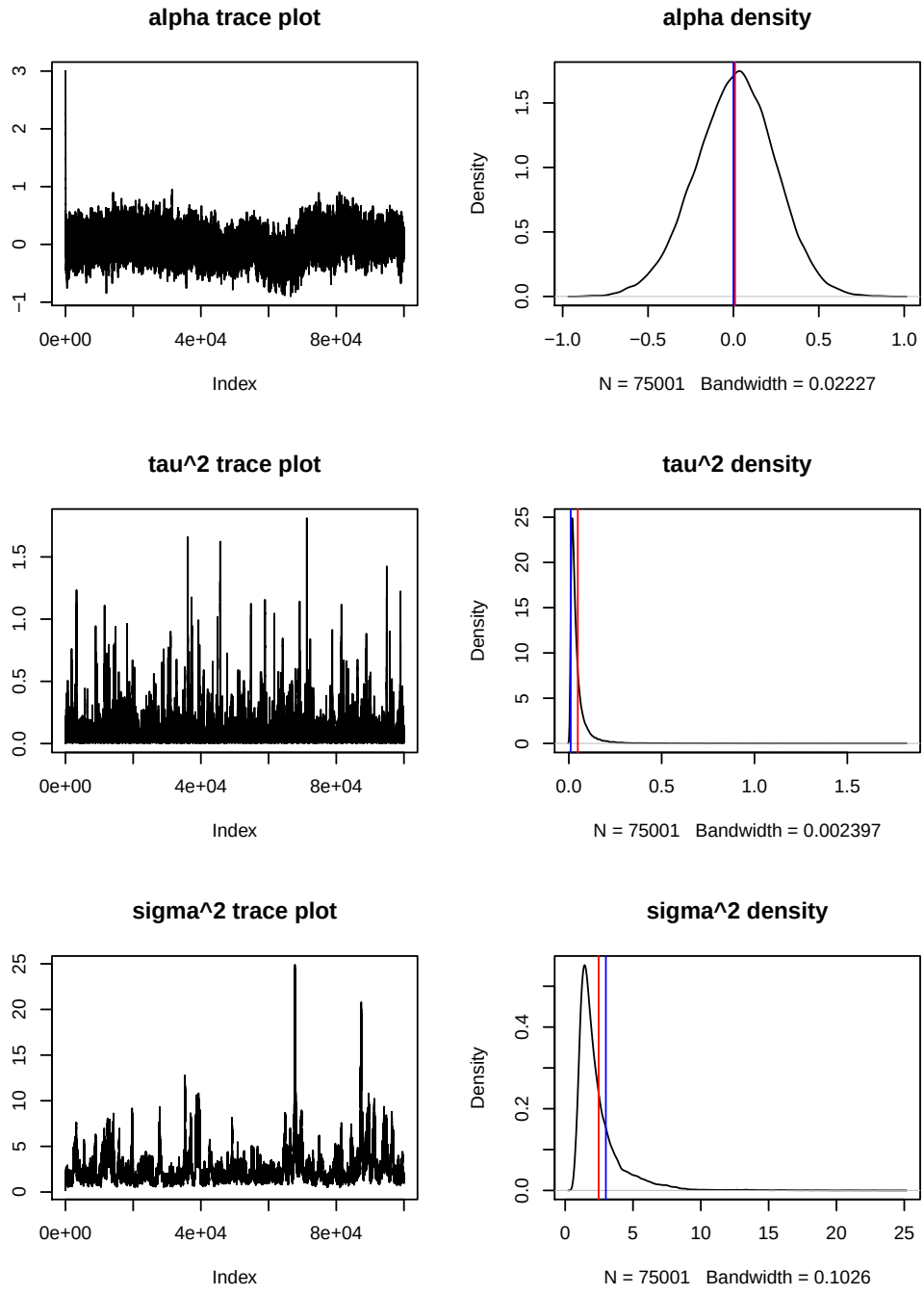


Figure 2.10: Parameter plots for model fit on non-preferentially sampled data, but a model which allows for preference. Red lines show predicted values, blue lines show actual values.

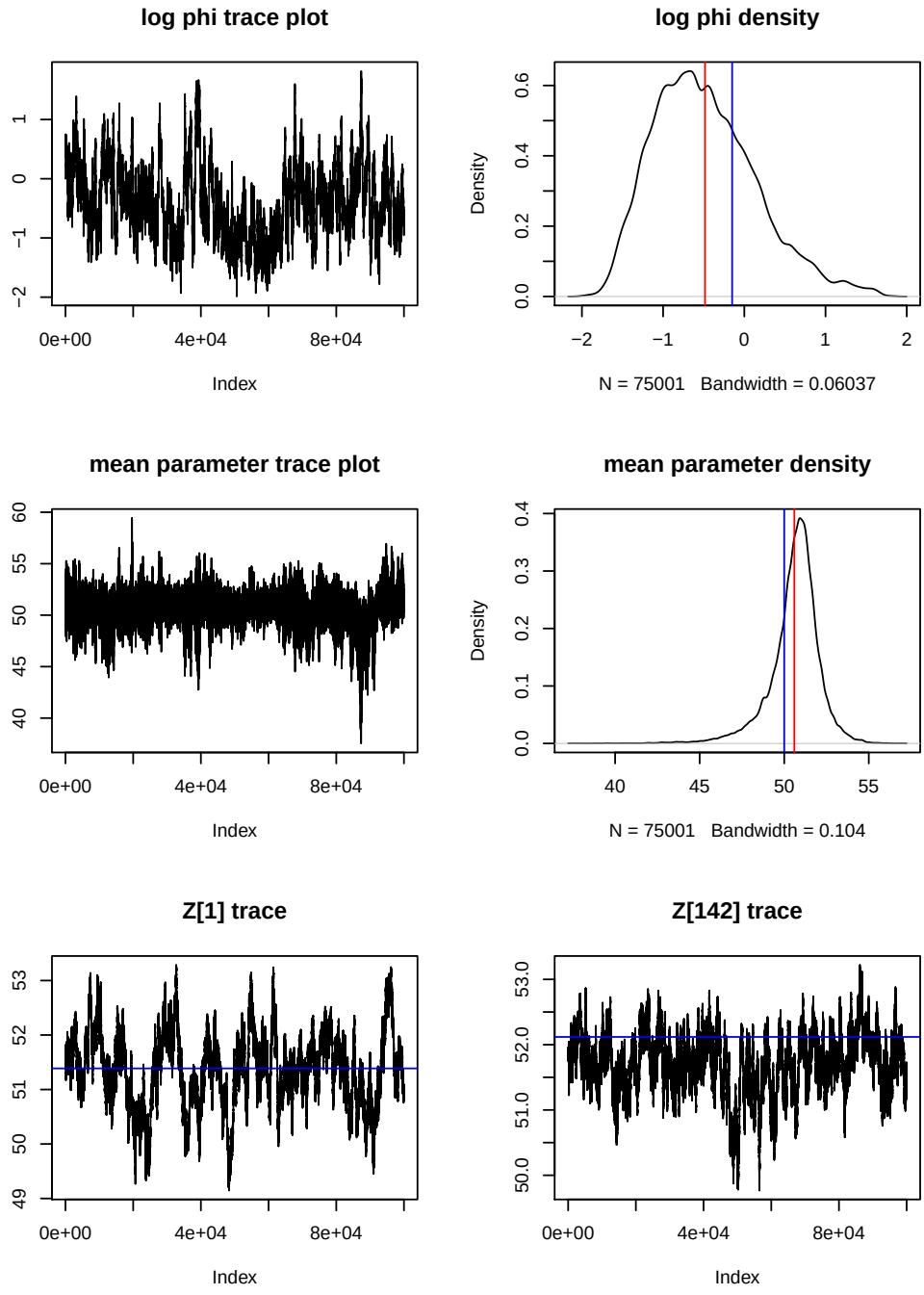


Figure 2.11: More parameter and Z value, plots for model fit on non-preferentially sampled data, but a model which allows for preference. Red lines show predicted values, blue lines show actual values.

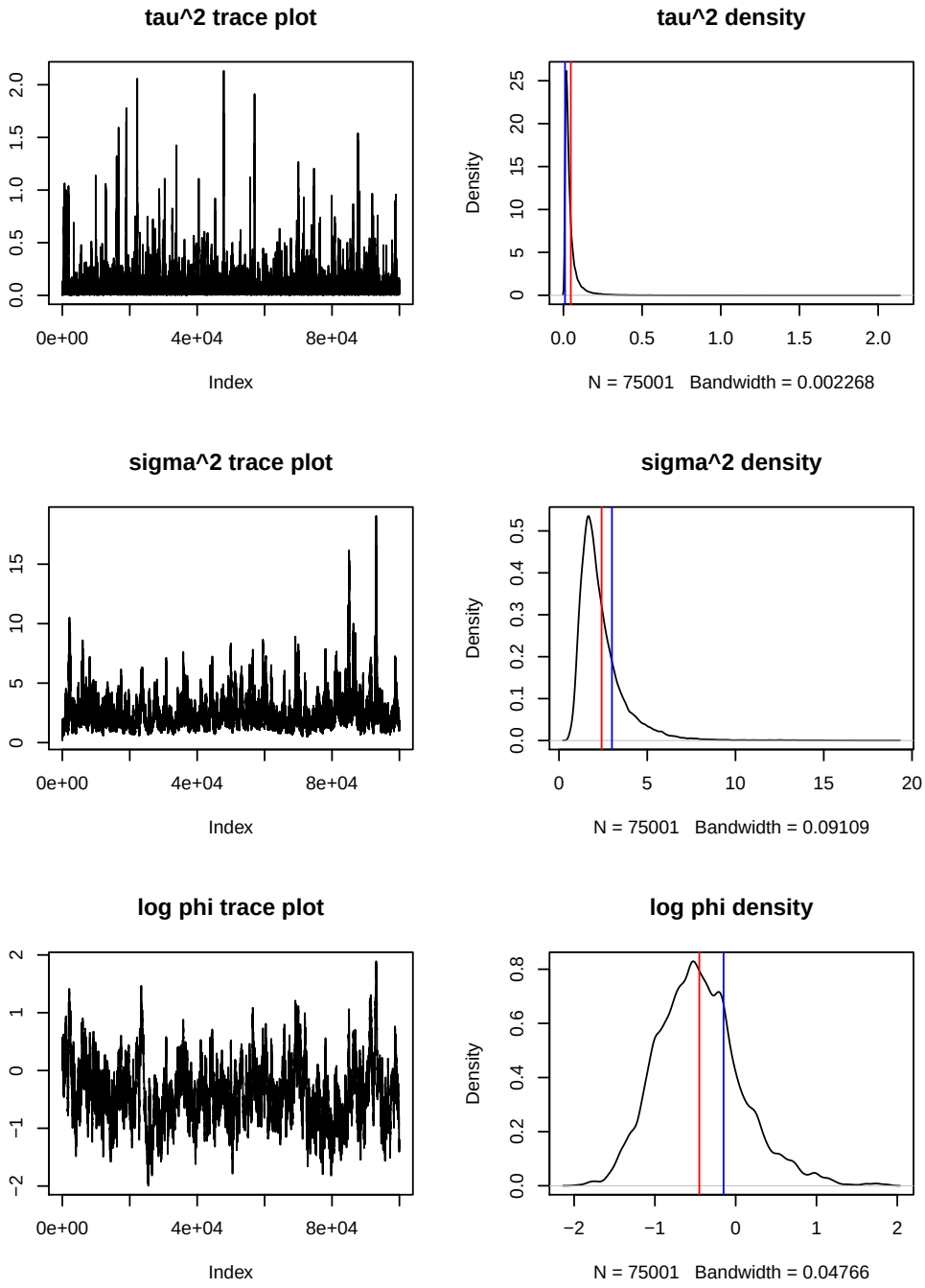


Figure 2.12: Parameter plots for model fit on non-preferentially sampled data, and a model which does not allow for preference. Red lines show predicted values, blue lines show actual values.

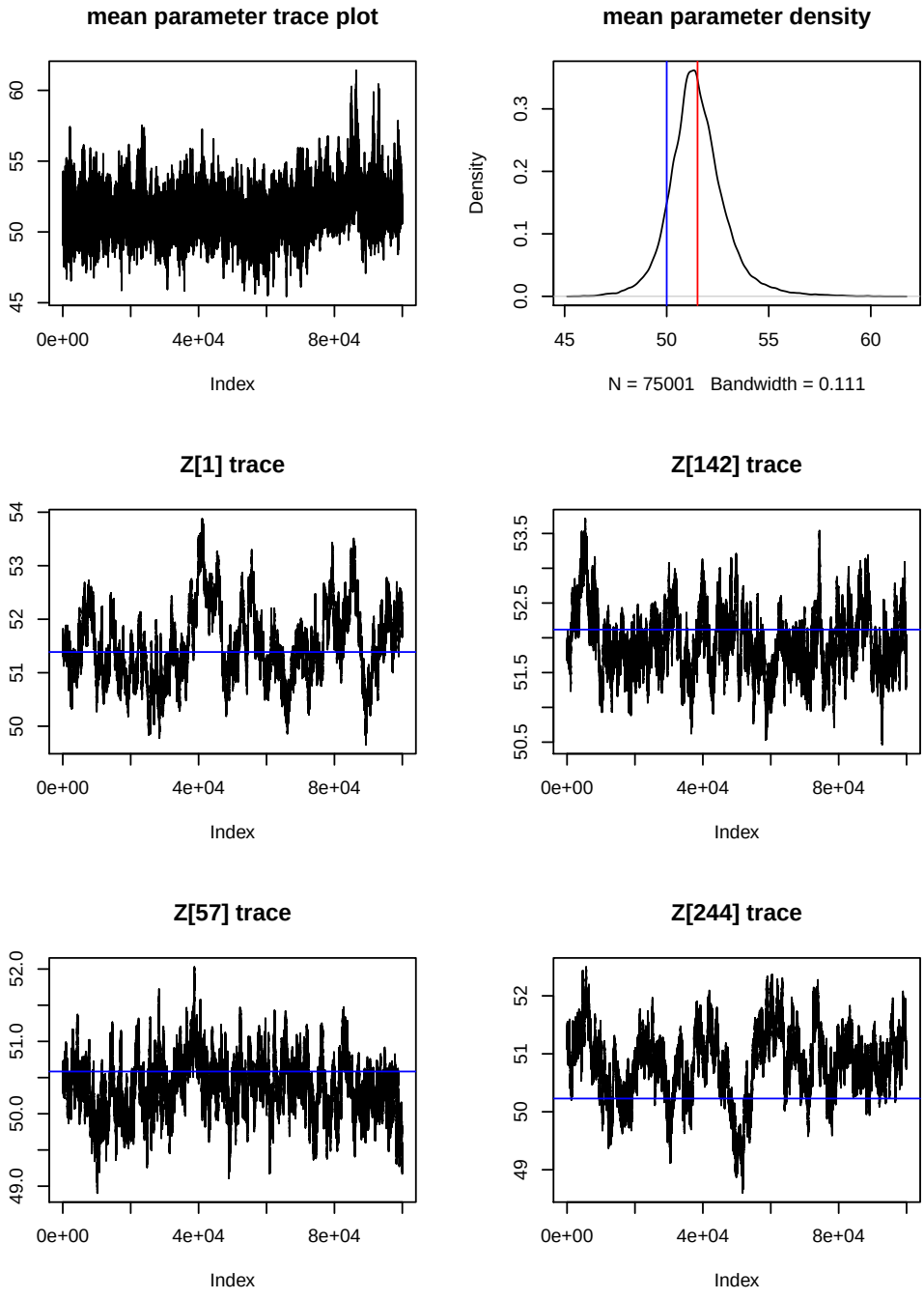


Figure 2.13: More parameter and Z value, plots for model fit on non-preferentially sampled data, and a model which does not allow for preference. Red lines show predicted values, blue lines show actual values.

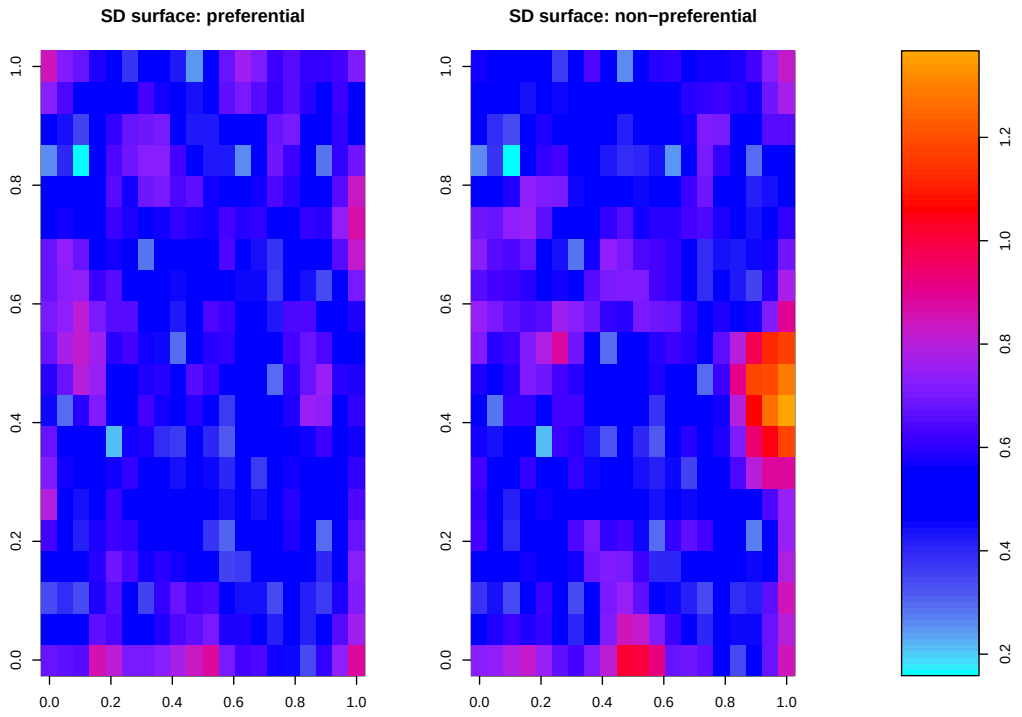


Figure 2.14: Standard deviation surfaces for Z, for both models fitted to the non-preferentially sampled data.

this is shown by the difference in the sum of squared errors: 196 where preferential sampling is taken into account, 446 where the preferential sampling is ignored. We can also see this in the comparative plots of the posterior mean surfaces (Figure 2.1) : the darker, lower valued regions have been overestimated by the non-preferential model. This is reflected in the fact that the mean squared error, and mean absolute error are exceptionally high (2.05 and 1.28) for the low Z values, when compared with both the corresponding values for which preferential sampling has been assumed, and the mean squared and absolute errors for the high Z values, which have not been predicted as badly by the non-preferential model. This can also be seen in Figure 2.7 when we see that for the lower values of Z there is a long 'tail' or non-adherence to the (0,1) line for in non preferential model, which is worse than that for the preferential model. We can see that, from Figure 2.1 for the preferential case, while no samples were taken in the upper left region, the model is able to identify, by detecting preferential sampling, that the values taken by the Gaussian random field in that region are likely to be lower. All the values of our metrics of comparison for Z are worse when we do not assume preferential sampling.

For this case in which we have preferentially sampled data, the standard deviation surface, shown in Figure 2.2 for the non-preferential model is similar but slightly larger, as one would expect when there is less available information. Similarly, we can see that, in both cases, the uncertainty is lowest round the sampled points, which is as we would expect.

We note that α has been slightly underestimated: this may be due in part to its prior mean having been set to 1: naturally it is possible that the preferential sampling could have been accounted for even

better with better estimation of α. The other parameters are all estimated reasonably well, with similar values for both the non-preferential and preferential models.

Finally, we also remark that in this situation, where there has not been any preferential sampling (i.e. the uniformly sampled data), using the 'wrong' preferential model has only had a small negative effect on prediction: α has been estimated as very close to zero: 0.00919 with standard deviation 0.234, and there is a difference of 8 in the sum of squared errors (much smaller than the difference of more than 200 between the two models for the preferential data). In fact, the predicted surfaces in Figure 2.8 appear to be almost identical, as is the level of adherence to the (0,1) line in Figure 2.9, and the standard deviation surfaces 2.14. The mean squared and absolute errors are similar for both models. This shows that in the situation in which one is not sure whether there has been preferential sampling or not, there is little danger that fitting a model which allows for preferential sampling will cause any catastrophic mis-estimation. In any case, if α were estimated to zero, it is likely that one might like to re-fit a non preferential model.

In these model fits to the preferential and non-preferential data, both φ and σ^2 are underestimated. As Z is well predicted, the problem here may be a matter of covariance parameter identifiability. Christensen et al. (2006) give some attention to this issue in the context of fitting spatial generalised linear models with exponential covariance functions (i.e. fitting random fields with such covariance functions, with linear predictors in the mean), and refer to Zhang (2004) who note that while σ^2 and φ are frequently poorly estimated, the estimates of their ratio are often more accurate. (Intuitively, this is because φ and σ^2 work in almost opposite directions: one makes close-by points more similar, the other makes them more different). Obviously our ability to estimate the values of these parameters accurately is determined by many factors, such as sample size, measurement error etc.. We encounter this issue of covariance parameter non-identifiability in further experiments, later in this thesis, fortunately, this has very little negative effect on the prediction of Z.

We have shown, by this illustrative example, both the problems that preferential sampling can cause, and that accounting for it within the hierarchical model for Z can enable these problems to be lessened. Obviously this is only one example: we shall demonstrate this further, with more complex utility functions in Chapter 5.

2.1.1 Experiments to show probabilities of over-estimation

In order to further demonstrate the effects of preferential sampling and the benefits of accounting for it, we repeat the experiment described in the previous section (with the same dimensions and parameters) but with 20 design realisations for a single Gaussian random field. We then attempt to recover this field (with 3000 MCMC iterations, the first 1000 of which are discarded as burn-in). Following the method of Gelfand et al. (2012) we then, for each of the $N = 400$ cell locations, find the probability that the associated Z value was overestimated. We can then find the mean value of this quantity over all N cells, (i.e. the mean overestimation probability) associated with each method. Where we used a model which accounted for preferential sampling, this mean probability was 0.656, as opposed to 0.808 for the model which did not account for preferential sampling. This shows that the model that does not account for preferential sampling is expected to overestimate the spatial process more often.

2.2 Bath air pollution data

Bath and North East Somerset have an air quality monitoring network, to monitor ambient levels of Nitrogen Dioxide. The network consists of 4 real time chemiluminescent monitors, and 109 diffusion tube monitors. We restrict our area of interest to just the city of Bath, between 161453 and 167820 in terms of Northings, and 372710 and 378022 in terms of Eastings. We consider the yearly averages for 2014. This gives us 38 data points around the city, as the numbers of operational monitors varied year by year.

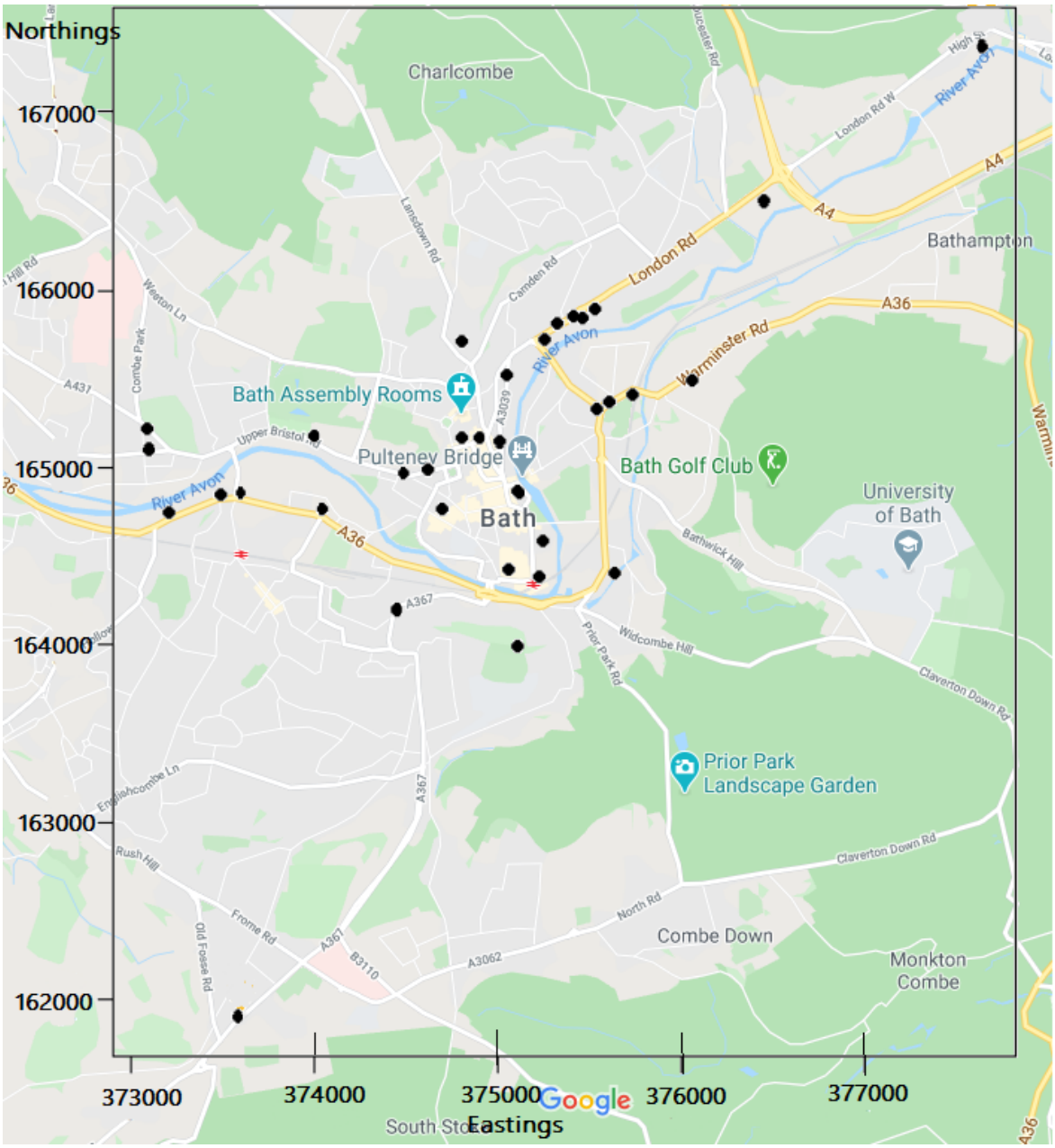


Figure 2.15: Map of the centre of Bath with the locations of the Nitrogen Dioxide monitors in 2014 superimposed. Map: ©2021 Google.

By looking at the map (Figure 2.15), and in particular noting that many of the points are situated

towards the centre of the city, or on roads, we have reason to believe that a preferential sampling model may be appropriate. We attempt to fit a model in which the points are assumed to have been selected via a multinomial utility function. A selection of other models for the preference, such as those including terms which rewarded good coverage of the city (as will be discussed in Chapter 3), and a model which included good coverage of the main roads using a non-euclidean distance-by-road metric were also considered and fitted, but gave very insignificant coverage preference parameters: thus we proceed with the multinomial utility.

We discretise the region in question into a 30×30 grid. We take distances in terms of Eastings and Northings, divided by 100000 (for computational reasons). In terms of prior distributions for the parameters, we have inverse gamma priors for τ^2, σ^2 and φ, with parameters $(1, 0.1)$ for φ, $(7, 70)$ for σ^2, $(2.1, 0.1)$ for τ^2 and an exponential covariance function. Non-informative normal priors with mean zero, and variance 20 were used for the strength of preference parameter α. We select an optimal sub-discretisation for the utility using the Deviance information criteria, using the methods as described in Chapter 6. We fit models, with 30×30, 15×15 10×10, 6×6 and 5×5 discretisations, which gave the lowest value for the original 30×30 grid for the utility, which we select. The model is fitted using MCMC with Metropolis Adjusted Langevin (MALA) (Roberts and Rosenthal (1998)) updates for the Gaussian random field Z, as will be described in Chapter 5. We sample 10000 times, with the first 5000 samples discarded as burn-in. For comparison we also include the case in which preferential sampling has not been assumed. Results are displayed in Figures 2.16, 2.18 and Table 2.3.

Results:

Parameter	Non-preferential mean (s.d.)	Preferential mean (s.d.)
α	NA	0.174 (0.0274)
σ^2	126.2 (35.8)	94.1 (17.0)
$\log(\varphi)$	-4.57 (0.281)	-4.79 (0.189)
τ^2	0.106 (0.254)	0.105 (0.168)
θ	34.7 (1.21)	35.2 (0.791)

Table 2.3: Predicted parameter values from the two models: preferential and non-preferential, for the Bath Nitrogen Dioxide Pollution data.

For this example, we have detected that preferential sampling is at work as shown by a positive predicted value of the strength of preference parameter α. Accounting for this has had a knock-on effect on the prediction of the underlying field, Z: in the preferential case, predicted values corresponding to the sampled cells have mean value of Z as 49.5, while the unsampled ones have a mean of 37.9. Conversely, in the non-preferential case these values are 47.8 and 38.9. This shows that when preferential sampling is assumed, on average, the unsampled cells and regions, i.e. not those close to the city centre, are estimated to have lower values of Z than in the non-preferential case. This is most clearly demonstrated in the the high-valued region in the North East of the region in question: in the preferential case, these high values are more restricted to the area directly around the monitor, whereas for the non preferential case, these high values are assumed to extend further, as less information is available to

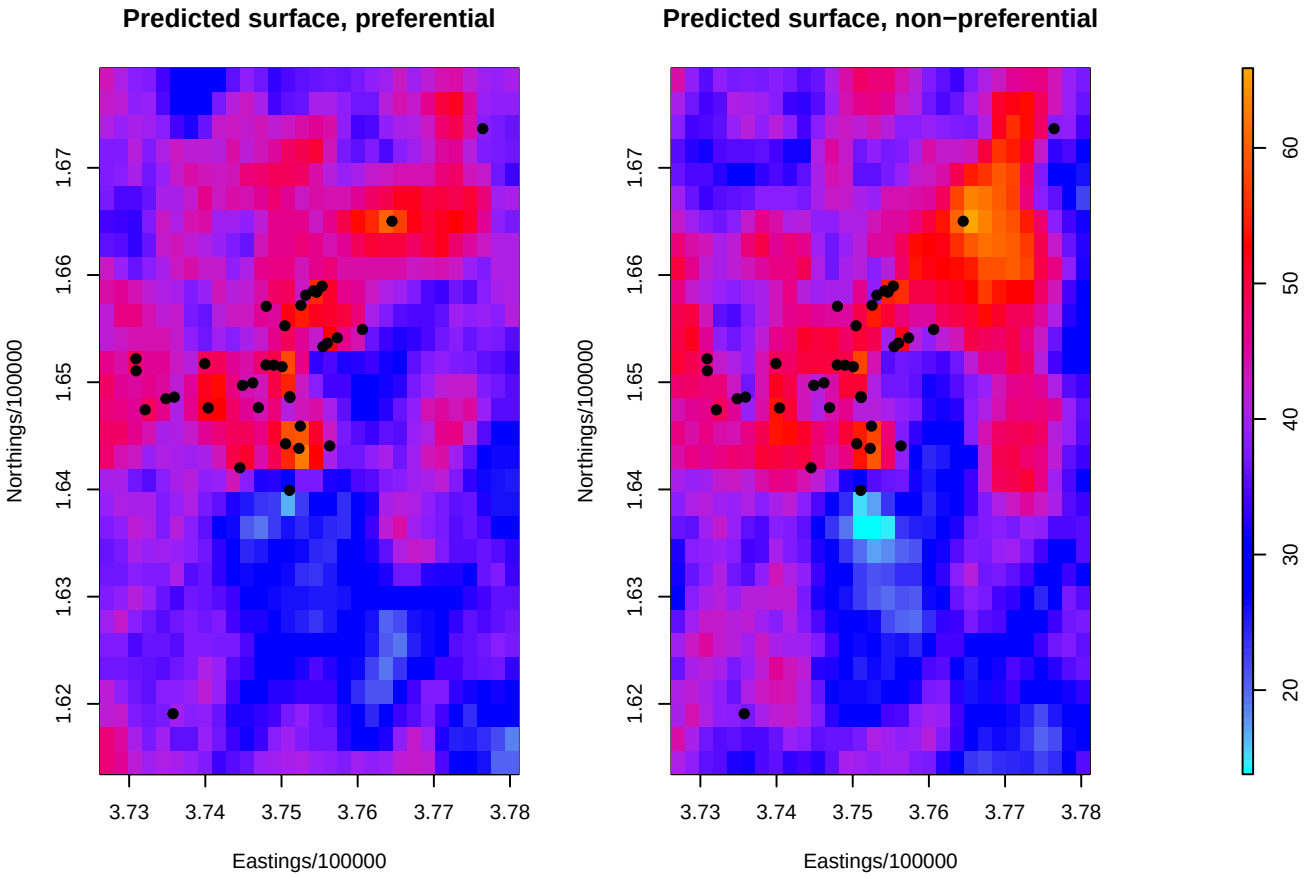


Figure 2.16: Comparison of Gaussian random fields for the predicted Nitrogen Dioxide concentrations for the city of Bath, 2014. The left hand image shows the results of the modelling when preferential sample is assumed, the right when it is not.

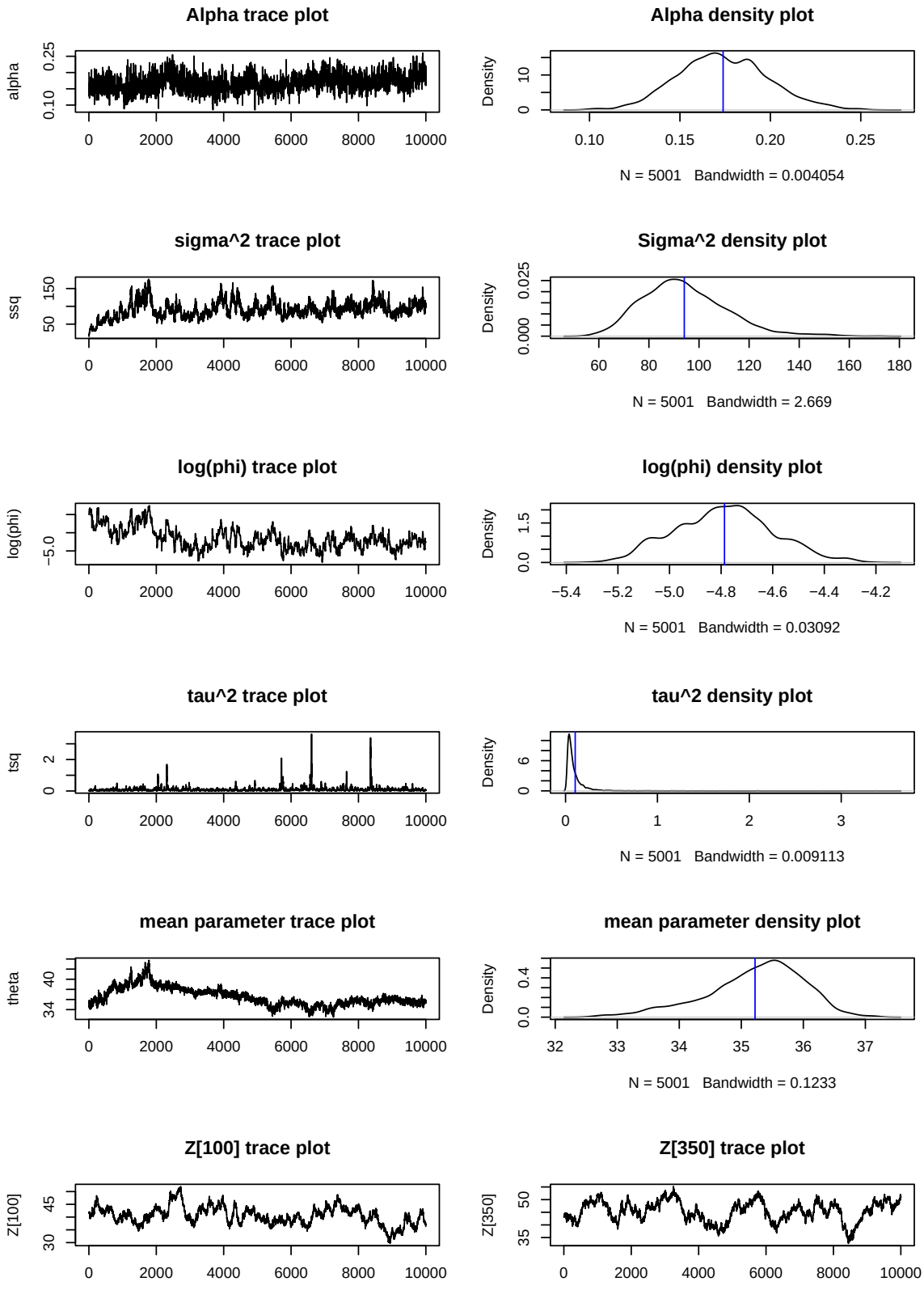


Figure 2.17: Parameters for preferential case, Bath Nitrogen Dioxide Pollution data. Blue lines indicate predicted (mean) parameter values.

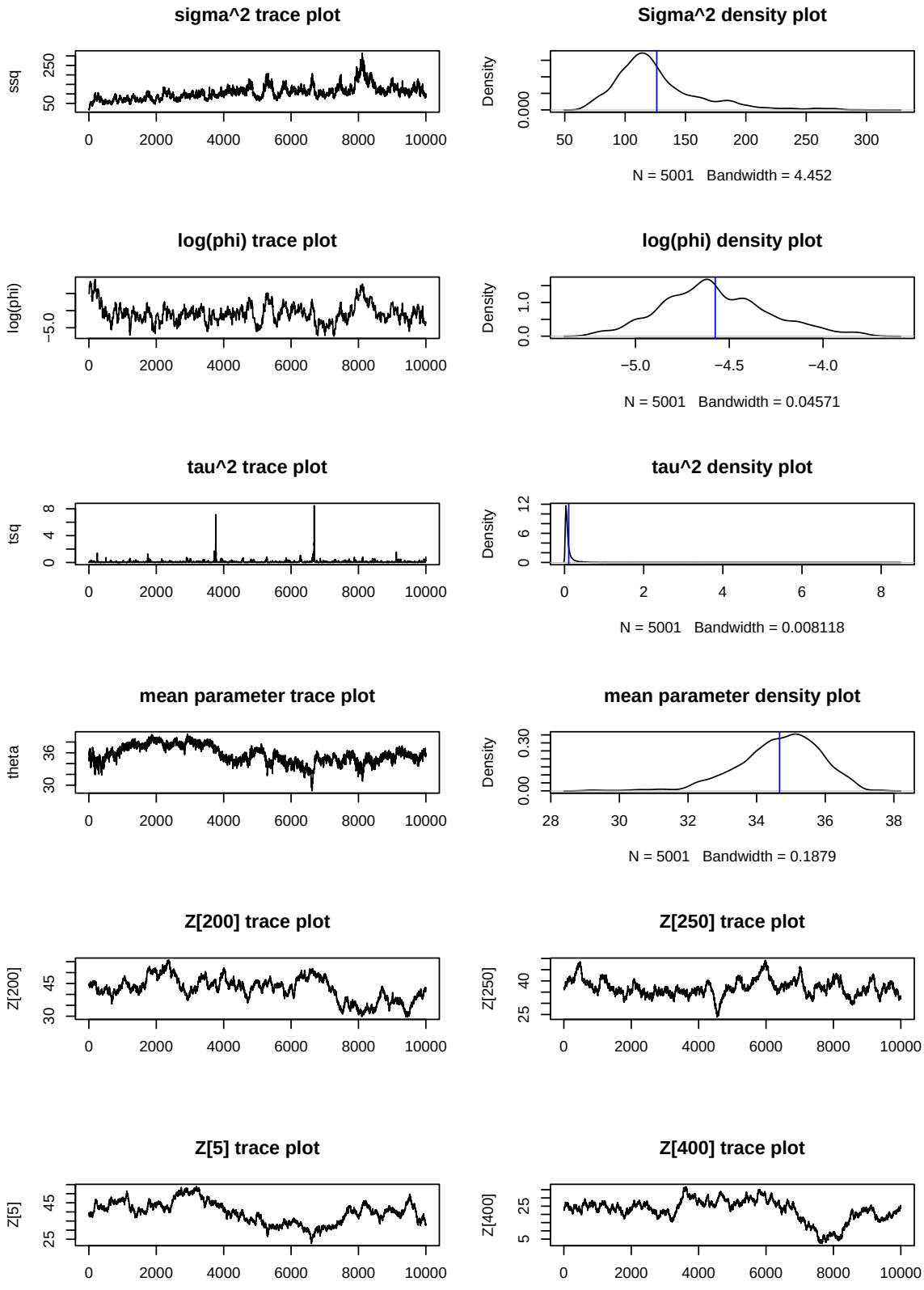


Figure 2.18: Parameters for non-preferential model fit, Bath Nitrogen Dioxide Pollution data. Blue lines indicate predicted (mean) parameter values.

suggest that the particularly high values are linked to monitor placement, and as we get further from the monitor, values are likely to decrease. The preferential model predictions seem likely to reflect reality, as urban activity, such as traffic, and solid fuel burning, is responsible for much of the pollution. The model seems to be worse for predicting pollution on main roads away from the main cluster of monitors around the city centre, for example, three monitors in a diagonal line in the North-East lie on the same main road, and so, intuitively, the correlation between those points, and the points along that road should be higher, possibly using a non-Euclidean distance metric. For readers familiar with the city of Bath, the monitor at the bottom of the central cluster is situated in Alexandra Park, atop a cliff separating it from the busy city centre: it is thus somewhat unsurprising that the lowest values should be recorded there. Future investigations could include altitude as an informative covariate.

2.2.1 Estimating preference and the number of monitors

In order to account for preferential sampling properly, we clearly need to be able to estimate the level of preference, α, well. The accuracy with which we are able to do this is dependent on several factors, such as the strength of preference and number of monitors we need enough monitors in lower valued areas, rather than having them entirely clustered in high valued areas, in order to establish a 'gradient' of monitor-concentration against measured values.

We demonstrate this with the following experiment: we generate a 10×10 Gaussian random field on the unit square with parameters $\theta = 50$, $\sigma^2 = 1$, $\tau^2 = 0.01$, $\log(\varphi) = -0.15$, and take n samples at irregular locations, selected according to a multinomial utility function (2.1), with the values of the Gaussian random field at those locations given mean zero Gaussian noise with variance τ^2. We then re-fit the model (as in Section 2.1) via MCMC with 5000 samples, with the first 1000 discarded as burn-in and find the predicted α value $\hat{\alpha}$. For each of the pairs of values of $n \in \{5, 10, 20, 25, 30, 40, 50, 60\}$ and $\alpha \in \{0.5, 1, 2, 3, 5\}$ we repeat this experiment 50 times, in order to find the mean squared prediction error in $\hat{\alpha}$, for different numbers of samples and strengths of preference. Results are shown in Figure 2.19, demonstrating how for higher strengths of preference a greater sample size is needed to estimate α with the same level of error. In the case of $\alpha = 5$ we can see that α requires more samples to estimate with the same accuracy. This is largely due to the fact that, in this case, with few monitors, they tend to be all clustered in a very small number of cells.

While the best strategy for determining whether preferential sampling may be detected would be simply to fit the model, allowing for preferential sampling, and seeing if the strength of preference parameters are positively estimated, one quick diagnostic indicator of whether one has sufficient information to model preferential sampling would be to plot the measured values against the distance to the nearest neighbouring monitor, in order to assess whether a link between higher monitor concentration and higher measured values might be established. Obviously the relevance of such a procedure requires that the scale of the variation of Z be comparable with the distances between monitors: it is of no use if all the monitors are placed in isolated high valued regions far from one another, with low valued regions in between. Similarly, where preference may not be estimated effectively from one set of data alone, where possible, the level of preference within it may be established using other data sets or prior information, as will be discussed in Chapter 7.

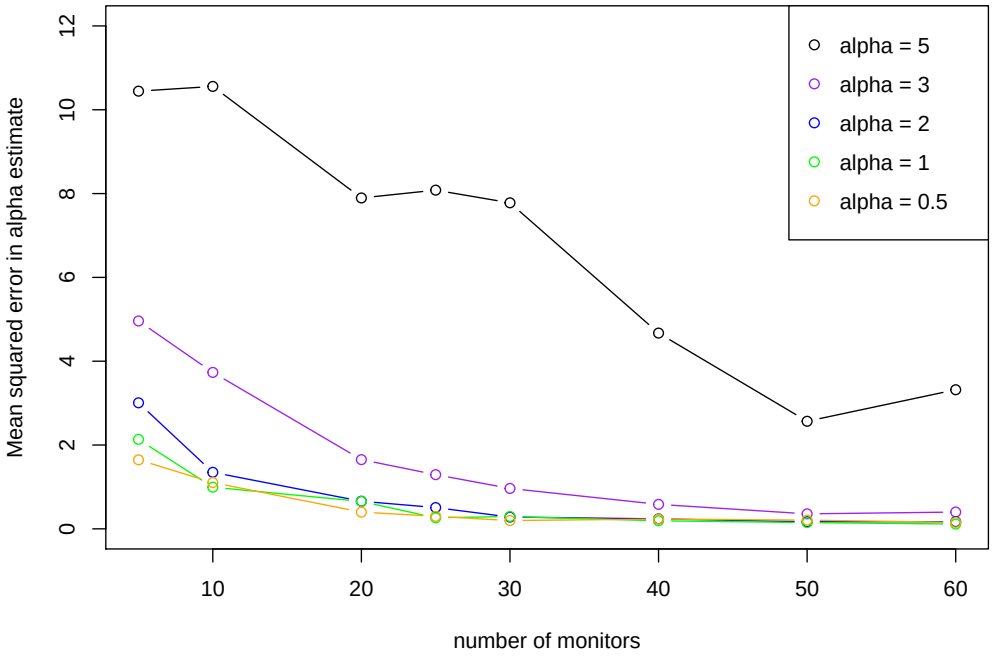


Figure 2.19: Mean squared error in the estimation of α for different values of n and α, calculated over 50 'fits' of the multinomial utility model. This demonstrates the need for larger sample sizes where there is a larger strength of preference, in order that a gradient may be established between monitor concentration and measured values.

Chapter 3

Non-multinomial utility functions

In this chapter we shall consider the selection of utility functions to define a 'whole-design utility' by which we may model the preferences of an experimenter in relation to the design, without the constraint of independence of sampling locations. The choice of an appropriate utility function will always depend on the specific situation, and any knowledge about the possible sampling process that has gone on. Clearly it is most important to consider the aspects of an experimenter's possible preferences which may have had to have been balanced against, or would have overinflated, the preference for high or low values, as these are the preferences which, if ignored, will lead to mis-estimation of the level of preference. With this in mind, we shall give particular focus to utility functions that model preference for high values, balanced with a possible preference for good coverage of the spatial region in question. Nonetheless, the methods presented (and the considerations made) may be easily adapted to a much wider range of possible preferences and corresponding utilities.

In addition to considering what the experimenter's actual intentions were, there are several considerations that must be made in this process of utility function selection. Firstly, in terms of how the preference is modelled, any candidate function must be able to encapsulate variable strengths of preference (including no preference), via 'strength of preference' parameters. This means, in practice, that higher values of these parameters should correspond to designs that exhibit more extreme levels of preference. This ensures the detectability of preferences, and, as we shall see, the selection of functions that enable this is a non-trivial problem. Finally, if in situations in which obtaining results very quickly is of importance, we might consider prioritising utility functions which display useful mathematical properties which allow for their normalising constant to be estimated more easily. More on these will be discussed in Chapter 4.

We begin this chapter with a consideration of some initially appealing functions which appear to encapsulate preferences in a very simple way, and whose normalising constant has a closed-form expression, which will motivate discussions of appropriate functional forms of utility functions. Next, we will focus our considerations on candidate functions for modelling balanced preferences for space-filling and high value preference, before considering some possible future directions in this area.

3.1 Polynomial-in-D utility functions

We begin our consideration of possible utility functions with functions which are polynomial in the elements of D. We do so as, not only are they intuitive for encapsulating preference, in that the 'rewards' for sampling certain designs may simply be the sum of the values of the process at them, they present the very attractive property of having easily calculable normalising constants. We define, for example a utility function of D-order 2 as follows:

Utility function of D-order 2:

$$U(D, Z; \alpha) = \sum_{i=1}^{N} f_i(Z; \alpha) d_i + \sum_{i=1}^{N} \sum_{j=1}^{N} g_{ij}(Z; \alpha) d_i d_j + C(Z; \alpha), \tag{3.1}$$

where each $f_i(Z; \alpha)$ and $g_{ij}(Z; \alpha)$, and $C(Z; \alpha)$ is an arbitrary function of Z.

Such utility functions work by the first, linear-in-D, term prioritising cells corresponding to higher values of $f_i(Z; \alpha)$ while the second order $d_i d_j$ terms control the pairwise interactions between the benefits of placing monitors in particular grid cells, for example, the gains of putting a monitor in a particular grid cell may be lessened if there is already a monitor in a neighbouring cell. Such functions appear to be very attractive due to the ease with which the normalising constant may be calculated exactly. Summing such utility functions over the space of all possible designs is straightforward as the computational burden may be greatly reduced by simply changing the order of the summation. We shall show this via the following proposition.

Proposition: Utility functions of the form 3.1 may be summed over the space of designs $\mathbb{D}$ as follows:

$$\begin{aligned}\sum_{D \in \mathbb{D}} U(D, Z; \alpha) = C(Z; \alpha) \binom{N+n-1}{n} + \sum_{i=1}^{N} f_i(Z; \alpha) \sum_{a=0}^{n} a \binom{N+n-(a+2)}{n-a} \\ + \sum_{i=1, i \neq j}^{N} \sum_{j=1}^{N} g_{ij}(Z; \alpha) \sum_{b=0}^{n} \sum_{a=0}^{n-b} ab \binom{N+n-(a+b+3)}{n-(a+b)} \\ + \sum_{i=1}^{N} g_{ii}(Z; \alpha) \sum_{a=0}^{n} a^2 \binom{N+n-(a+2)}{n-a},\end{aligned} \tag{3.2}$$

Proof:
We will make extensive use of the fact that there are $\binom{N+n-1}{n}$ possible different designs when we have N possible sites and n monitors to put in them. First we look at the constant and first order terms:

$$\sum_{D \in \mathbb{D}} \sum_{i=1}^{N} (f_i(Z; \alpha) d_i + C(Z; \alpha)) = \sum_{D \in \mathbb{D}} C(Z; \alpha) + \sum_{i=1}^{N} f_i(Z; \alpha) \sum_{D \in \mathbb{D}} d_i.$$

Clearly $\sum_{D\in\mathbb{D}} C(Z;\alpha) = C(Z;\alpha)\binom{N+n-1}{n}$ simply by counting the sites. Furthermore we have

$$\sum_{D\in\mathbb{D}} d_i = \sum_{a=0}^{n} a \sum_{D\in\mathbb{D}} 1(d_i = a) = \sum_{a=0}^{n} a\binom{N+n-(a+2)}{n-a},$$

as the number of monitors in site i can vary from 0 to n. When the number of monitors is a, then there are $n-a$ monitors left to position over a choice of $N-1$ sites, thus $d_i = a$ appears in $\binom{N+n-(a+2)}{n-a}$ designs. This means we have

$$\sum_{D\in\mathbb{D}}\sum_{i=1}^{N}(f_i(Z;\alpha)d_i + C(Z;\alpha)) = C(Z;\alpha)\binom{N+n-1}{n} + \sum_{i=1}^{N} f_i(Z;\alpha)\sum_{a=0}^{n} a\binom{N+n-(a+2)}{n-a}.$$

The second order terms may be dealt with in a similar way:

$$\sum_{D\in\mathbb{D}}\sum_{i=1}^{N}\sum_{j=1}^{N} g_{ij}(Z;\alpha)d_i d_j = \sum_{i=1}^{N}\sum_{j=1}^{N} g_{ij}(Z;\alpha)\sum_{D\in\mathbb{D}} d_i d_j.$$

First we consider $\sum_{D\in\mathbb{D}} d_i d_j$ when $i \neq j$. When d_i may take any value between 0 and n, d_j may then take any value between 0 and $n - d_i$. The combination $d_i = a$ and $d_j = b$ (within these bounds) then occurs in $\binom{N+n-(a+b+3)}{n-a-b}$ designs as there are $n-(a+b)$ monitors left to position, over a possible $N-2$ sites. This leads to, for $i \neq j$

$$\sum_{D\in\mathbb{D}}\sum_{i=1}^{N}\sum_{j=1}^{N} g_{ij}(Z;\alpha)d_i d_j = \sum_{i=1}^{N}\sum_{j=1}^{N} g_{ij}(Z;\alpha)\sum_{a=0}^{n}\sum_{b=0}^{n-a} ab\binom{N+n-(a+b+3)}{n-(a+b)}.$$

Finally, for $i = j$ we have

$$\sum_{D\in\mathbb{D}}\sum_{i=1}^{N} g_{ii}(Z;\alpha)(d_i)^2 = \sum_{i=1}^{N} g_{ii}(Z;\alpha)\sum_{D\in\mathbb{D}}(d_i)^2,$$

so we consider $\sum_{D\in\mathbb{D}}(d_i)^2$. d_i can take any integer value from 0 to n with $d_i = a$ occurring in $\binom{N+n-(a+2)}{n-a}$ designs as there are $n-a$ monitors left to place, across $N-1$ sites. This leads to

$$\sum_{D\in\mathbb{D}}\sum_{i=1}^{N} g_{ii}(Z;\alpha)(d_i)^2 = \sum_{i=1}^{N} g_{ii}(Z;\alpha)\sum_{a=0}^{n} a^2\binom{N+n-(a+2)}{n-a}.$$

Collecting all terms together leads to the final formula (3.2). The same logic may be easily applied to third and higher order utilities etc. □

This formula would be particularly useful when the normalising constants $K(Z;\alpha) = \sum_{D\in\mathbb{D}} U(D,Z;\alpha)$ are calculated repeatedly, as in an MCMC context, as they may be re-written as

$$K(Z;\alpha) = C(Z;\alpha)A + B\sum_{i=1}^{N} f_i(Z;\alpha) + E\sum_{i=1,i\neq j}^{N}\sum_{j=1}^{N} g_{ij}(Z;\alpha) + F\sum_{i=1}^{N} g_{ii}(Z;\alpha),$$

where A, B, E and F only depend on n and N and may therefore be calculated only once, regardless of new values of Z. This leads to a computational burden similar to that of one utility function evaluation (or smaller), rather than the $\binom{N+n-1}{n}$ required to enumerate all of them.

3.1.1 Linear utility functions and preferential sampling

While it is easy to find the normalising constants associated with utility functions that are polynomial in D, we must consider whether they are of practical value in describing preferential sampling: whether they may actually be used to describe a preference for sampling from high-valued sites. We can put some sensible requirements on these functions, or the terms within them which describe the possible strengths of preference. Say we consider functions $U(D, Z; \alpha) = \sum_{i=1}^{N} f(z_i; \alpha) d_i$, where α is a scalar which adjusts the strength of preference for higher values of Z.

1. $f(z_i, \alpha)$ must be increasing in both terms, so that larger z_is are favoured, and the strength of preference for them may be adjusted.

2. $\frac{U(D_1, Z; \alpha)}{U(D_2, Z; \alpha)}$ must be increasing in α where D_1 is a design that is preferable to D_2, given Z. i.e. α controls the strength of preference.

3. $f(z_i; \alpha)$ must always be positive.

The second and third conditions are satisfied by exponential functions, and the first condition might suggest, for example, $f(z_i, \alpha) = \exp(\alpha \exp(z_i))$ where the double exponential ensures that higher values α always increase $f(z_i, \alpha)$, even when z_i is negative. In such functions, a large increase in α leads to the term relating to the highest $z_i = z_M$, with count index d_M, dominating the sum $\sum_{i=1}^{N} f(z_i, \alpha) d_i$, and, as $\alpha \to \infty$ the utility of a design becomes proportional to d_M, while the choices made about the other sites rapidly become irrelevant. There are $\binom{N+n-k-2}{n-k}$ designs involving k sites at the location of z_M, so, when we have

$$P(D|Z, \alpha) \;\propto\; U(Z, D; \alpha) \;\propto\; d_M,$$

as we do as $\alpha \to \infty$, the probability of selecting any one of them will be

$$P(d_M = k) = \frac{k\binom{N+n-k-2}{n-k}}{\sum_{j=0}^{n} j\binom{N+n-j-2}{n-j}}.$$

This means that

$$\frac{P(d_M = k)}{P(d_M = k+1)} = \frac{(N+n-k)k}{(n-k+2)(k+1)},$$

which, provided $N > n + 3$, will be greater than 1 for $k \geqslant 1$. This means that the designs most likely to be produced will contain one, or a very small number, of samples taken at the site corresponding to z_M, with the rest scattered uniformly across the other sites, as the contribution to the utility made by the other sites is largely irrelevant. This calls into question how realistic these functions are to describe preferential sampling, as increasing α arbitrarily cannot lead to probability distributions that give large weight to clusters of many monitors in the same site, i.e. a large d_M. Secondly, even if the designer wished to have designs which had at least one of the highest-valued site, but then had little preference

with regard to the other sites, they would have to know with certainty which site would have the highest value for this utility function to make sense. Furthermore, it would be very difficult to estimate the parameters of this model, as the distribution is relatively flat, no matter the value of α: design draws from the utility with high values of α are generally indistinguishable from those with $\alpha = 0$. This leads us to the conclusion that these are not sensible choices of utility function.

3.2 Space-filling utilities

A reasonable preference an experimenter may have is for a sampling design that has the monitors spread out to provide even coverage of a region. The beneficial properties of such designs are discussed in Nychka et al. (1997). Likewise, examples of such preferences are discussed in the literature, for example Fernández et al. (2005), with reference to the 'Heavy metals in European mosses project', propose that a gridded sampling design is favourable when considering objectives such as mapping concentrations and determining the effects of emissions sources. Likewise, Gelfand et al. (2012) note that it would be unusual for sites to be chosen randomly according to some intensity function based on the underlying surface, with geometric and space-filling designs often begin preferable. We consider the fact that a preference for high values may be balanced with such a preference for good coverage. It is important to take this into account as, if ignored, the high value preference may go undetected. For example, a cluster of points may occur with reasonable likelihood in a high-valued area under the assumption of uniformly-sampled points, but not under the assumption that the experimenter's default preference may be to spread the monitors evenly.

We seek a utility function by which we are able to model this mixed preference. Given the failure of polynomial functions of Section 3.1 to encapsulate either preferential sampling or interactions between sampling sites, and in order to define a model flexible enough to model the situation in which there is no space-filling but still a high value preference, we choose a function that reduces to the multinomial utility when the preference for space-filling, which may be controlled by a strength of preference parameter β, is zero. More specifically, we seek a function $g(D)$ such that the utility function

$$U(Z, D; \alpha, \beta) = \frac{n!}{\prod_{i=1}^{N} d_i!} \exp\left(\alpha \sum_{i=1}^{N} z_i d_i + \beta g(D)\right), \tag{3.3}$$

is able to encapsulate this mixed preference. This preference for space-filling can generally be broken down into a preference for two things: that no substantially large areas of the region in question be unsampled, or, in other words, nowhere is too far from a sampling site, and that the sampling sites themselves are not clumped too close together. We explore a selection of functions by which these preferences can be expressed.

There are several commonly used criteria that may be maximised in order to give optimal space-filling designs, some of which are discussed in Pronzato and Müller (2012). However, it is worth noting that our objectives are subtly different: instead of seeking the designs which optimally fill the space, we seek a function that assigns some sensible grading of monitor spread to designs that, due to a concurrent preference for high values, will inevitably be suboptimal with respect to space-filling. In other words,

where any one design is strictly preferable to another, in terms of the extent to which the space is evenly filled, we require a criterion which takes a value that is strictly greater for the preferable design, even when neither is optimal in terms of space-filling. It is important to stress the difference between this objective and the objective of a criterion which simply assigns the highest value to the optimal design in terms of even space-filling, but does not necessarily provide such a ranking to space-filling-suboptimal designs. We consider several possible functions and their suitability.

3.2.1 Minimax and maximin criteria

The first criteria we consider are the minimax and maximin criteria. The first of these prioritises closeness of every point to a monitor, while the second prioritises the distances between the monitors. The key difference between the designs produced by optimising these criteria is that the second favours the positioning of monitors close to the boundary of the region while the first does not.

Minimax criterion

The minimax criterion takes the value of the maximum distance between any cell j and its nearest monitor:

$$g_{mM}(D) = -\max_{j\in(1,\ldots,N)}\left(\min_{i:d_i>0} M_{ij}\right), \tag{3.4}$$

Where M is the pairwise distance matrix between the focus points of all cells within the region in question. This criterion is to be minimised, hence the minus sign before the function. When we consider the designs for which this criterion is relatively high, but not not maximised, it becomes clear that this function does not provide a reasonable ranking of designs. Figure 3.1, showing two designs that are equivalent in terms of the value of (3.4), taking a value of 0.19, demonstrates this: the only criterion by which designs are ranked is the maximum distance, not taking into account that such a criterion value can be attained by designs both with and without clusters of monitors. Furthermore, when used in a combination utility, there may be a situation in which there is an area of the region in which the process of interest is assumed to be so low that there would be no point in putting monitors there, or there is a large gap in points due to the mis-specification of the region boundaries, or geographical features, leading to a high value of this criterion. In this case the spacing between the remaining points, provided that a larger gap between them is not made, is irrelevant.

Maximin criterion

The Maximin criterion takes the value of the minimum distance between any monitor and its nearest neighbouring monitor.

$$g_{Mm}(D) = 1(\max_i d_i = 1)\min_{j:d_j=1}\left(\min_{k:d_k=1,k\neq j} M_{ij}\right) \tag{3.5}$$

This criterion is to be maximised. We see a similar, but opposite effect as with the minimax criterion, as shown by the leftmost two plots in Figure 3.2. Additionally, if there is at least one point in a cell, or points in neighbouring cells, the criterion takes the corresponding small value, and the spread of the

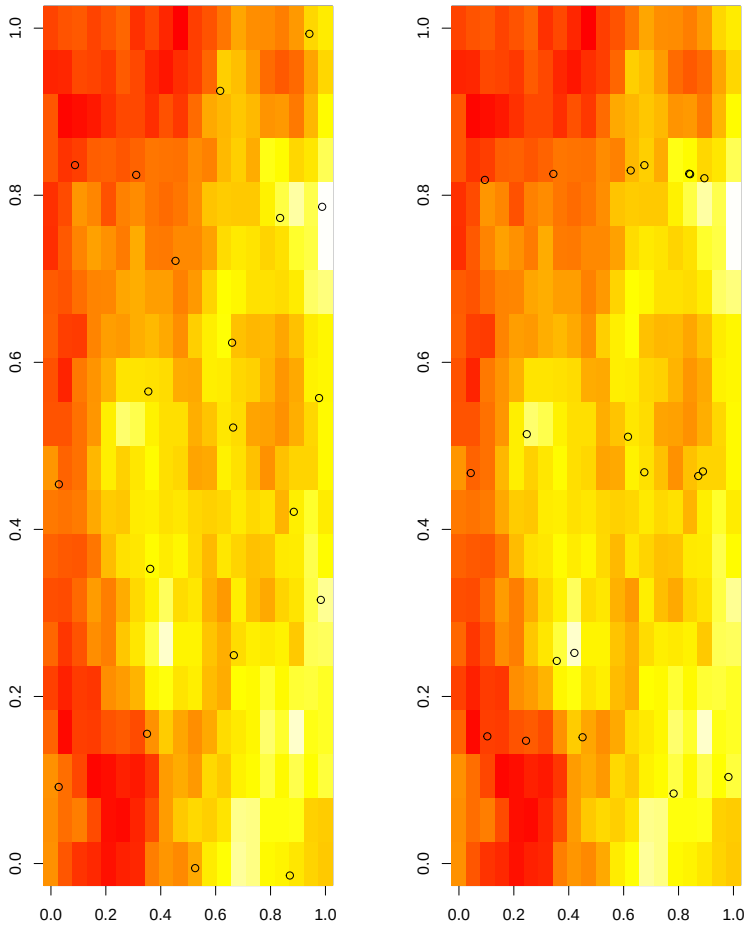


Figure 3.1: This figure shows two designs that are equivalent in terms of the maximum distance from any cell to its nearest sampled point, i.e. the minimax criterion given in Equation (3.4). While the designs here do not depend on the values of Z, the different colours along the yellow-red spectrum indicate Gaussian random field values from high to low.

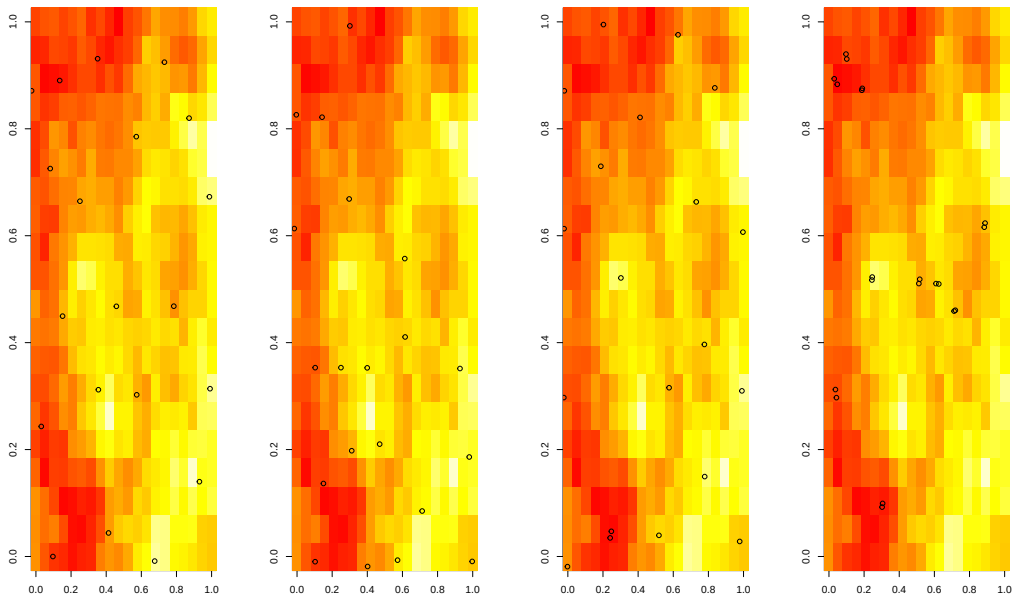


Figure 3.2: The leftmost two plots in this figure shows two designs that are equivalent in terms of their value of the maximin criterion: (3.5). Likewise, the rightmost two plots show two more designs that are equivalent in terms of the value of (3.5). This is because they both contain at least one cell containing more than one sampling site. While the designs here do not depend on the values of Z, the different colours along the yellow-red spectrum, included for clarity, indicate Gaussian random field values from high to low. Such designs are generated by Metropolis-Hastings point swapping, as in 4.

other points becomes irrelevant. For example, the rightmost two plots in Figure 3.2 show two designs for which this criterion is zero, as each has at least one cell with a repeated point. This makes this criterion particularly unsuitable for use in a combination utility, as high preference for certain areas makes the positioning of multiple monitors in one cell a reasonable possibility. Likewise, the distribution is rendered very sensitive to the size and location of cell boundaries.

Minimax-maximin combination criterion

We may also consider criteria in which a linear combination of the above minimax and maximin criteria are used, however, the fact that the maximin criterion is rendered useless in this instance by having more than one monitor in a cell leads us to conclude that this would be only a slight improvement.

Convergence of chains induced by minimax and maximin criteria

The nature of such criteria means that the Markov chain by which we can sample from the distribution of designs proportional to utility functions that incorporate them is slow to converge. This is due to the fact that for an increase to be made in the utility a point must be moved in such a way as to deal with the largest distance between any cell and its nearest monitor (i.e. the largest gap), meaning that the probability of moving in one step to a design of higher utility is very low, for a traditional Metropolis Hastings algorithm. Additionally, moves of points to large (but not largest) gaps between points, which would represent a considerable improvement in the space-filling (once the largest gaps had been dealt with), do not precipitate an increase in utility. This slow convergence is a problem as, in the wider prediction Markov chain used to fit the joint spatial and sampling process model (as in Chapter 5) we require design samples at each iteration, for the sampling of Z, α and β.

3.2.2 Pairwise distance and coverage functions

A key problem with the two criteria described by Equations (3.5) and (3.4) is their lack of dependence on every point in the design: once the criterion cannot go above a certain level for a reason such as a twice-sampled cell or a single large gap between monitors, the positioning of the other points makes no difference. In order to remedy this, we consider the following function, which depends on the pairwise distances between each monitor, and the distances between each cell and each monitor:

$$g_{PD}(D) = \sum_{j=1}^{N}\sum_{i=1}^{N} M_{ij}d_i d_j - \left(\frac{2n}{N}\right)\sum_{j=1}^{N}\sum_{i=1}^{N} M_{ij}d_i,$$

This function, via the second term, rewards a small sum of the distances between each cell and each monitor, while penalising small distances between monitors, via the first term. The $\frac{2n}{N}$ ratio is derived from the assumption that when $n = aN$ for some positive integer a, the function should take its maximum value when there are a monitors positioned in each cell. This function is much more dependent on the whole design, rather than maxima and minima, and draws from this design appear to represent mixed preferences to a reasonable level. However, due to the fact that for each cell i, the values $\sum_{j=1}^{N} M_{ij}$ are not equal (it generally takes smaller values in the centres of regions), the value of the criterion for

designs with single large clusters, even with otherwise well spread out points (which may occur with weak or moderate preferences for space-filling), is highly dependent on the location of the cluster within the region, giving higher values to designs with clusters in the centre, and lower values to designs with clusters in one corner of the region. Thus, a combination utility that employs this function is less likely to detect preferences for space-filling if the higher values are non-central in the region. For this reason we do not recommend the use of this function, and we shall, henceforth, not use it.

3.2.3 Mean distance functions

We now consider the functions that use as a criterion either the mean distance from each monitor to its nearest neighbouring monitor or the mean distance from each cell to its nearest monitor. We will refer to these as the mean-nearest-neighbour, and mean-coverage functions.

Mean-nearest-neighbour:

$$g_{NN}(D) = \begin{cases} 0 & \text{if } \nexists\ i \text{ s.t. } d_i = 1 \\ \dfrac{1}{n} \displaystyle\sum_{i:d_i=1} \left(\min_{j \neq i, d_j > 0} M_{ij} \right) & \text{otherwise.} \end{cases} \tag{3.6}$$

Mean-coverage:

$$g_{MC}(D) = -\frac{1}{N} \sum_{j=1}^{N} \left(\min_{i:d_i>0} M_{ij} \right). \tag{3.7}$$

These, while motivated by the minimax and maximin criteria, are far less sensitive to individual points and gaps between points, giving greater distinguishability to different designs. We construct an example using simulated designs to demonstrate the different extent to which different criteria distinguish between different designs. Out of 500000 unique designs of size $n = 30$ with $N = 625 = 25 \times 25$ possible cells, generated by simple random sampling with replacement, the minimax criterion $g_{mM}(D)$ (3.4) took 226 unique values, the maximin criterion $g_{Mm(D)}$ (3.5) 32, as opposed to 482571 for the mean-nearest-neighbour distance $g_{NN}(D)$ (3.6), and 500000 for the mean-coverage function $g_{MC}(D)$ (3.7). This demonstrates the greater ability of the mean-value criteria to distinguish between designs in terms of the goodness of their spatial coverage. Additionally, the dependence only on nearest neighbour distances, as opposed to all pairwise distances between monitors and points (as in the pairwise-distance function) avoids discrepancies between designs based on the locations of clusters in otherwise well-spaced designs.

We propose that it would be advisable to use a linear combination of the two mean distance criteria. When we are only interested in the maxima of these criteria, (aside from differences in how close to the boundary points are positioned), when one criterion is maximised, the other is generally also high. However, when we are only looking at the moderately high values of these criteria, this is not always the case. To illustrate this, Figure 3.3 displays four designs, the top two of which have the

same mean-nearest-neighbour distance, but are different in terms of their mean-coverage distances. The bottom row shows two designs that are equivalent in terms of their mean-coverage distance, but different in terms of their mean-nearest-neighbour distances. This use of a linear combination also ensures distinguishability between designs for which one of the criteria is high because it takes the value of the mean of one high and many low values to be averaged, or the mean of roughly equal values, which would indicate more even spacing. This motivates the question of what the coefficients of such a linear combination should be. Clearly, for the purposes of making designs with different values of each of the two criteria distinguishable, we only require that they both be positive. There seems to be little reasoning for specific values, however it should be acknowledged that giving more weight to the mean-nearest-neighbour distance criterion slightly increases the weighting to designs with monitors close to the boundary. Henceforth we shall use the 'combination utility'

$$U(Z, D; \alpha, \beta) = \frac{n!}{\prod_{i=1}^{N} d_i!} \exp(\alpha \sum_{i=1}^{N} z_i d_i + \beta g(D)) \tag{3.8}$$

with

$$g(D) = g_{MC}(D) + g_{NN}(D),$$

where $g_{MC}(D)$ and $g_{NN}(D)$ are defined as in Equations 3.6 and 3.7. This function effectively encapsulates mixed preferences. Designs on the unit square drawn from the implied distribution are shown in Figure 3.4.

It is important to note that the respective sizes of α and β associated with the relative importance placed on sampling large values compared with that of spreading monitors out varies according to the number of monitors n. While the 'sum' term scales roughly with n, the 'spread' term, based on shortest distances, scales roughly with $\frac{1}{\sqrt{n}}$. This, for the sum term is because it is made up of n αz_i values, whereas for the spread term comes from the fact that when we have n monitors scattered uniformly over a region of area A, we would expect there to be $\frac{n\pi r^2}{A}$ monitors in a circle of radius r around any point, so the shortest radius r of a circle round a point in which we would expect there to be exactly one other monitor would be roughly $\frac{\sqrt{A}}{\sqrt{\pi n}}$.

The discretisation also has relevance for the use of such a combination utility function. In order to take advantage of the properties of the function combination utility (3.8) which make it useful for modelling preferences for space-filling designs, a grid size should be selected such that $N > 2n$. This is because, once there is already a monitor in every other cell, increases in the value of the function do not necessarily coincide with a more even spread of the remaining monitors (i.e. an even spacing of the gaps between monitors). In fact, when $N < n$ the highest values of the function are achieved when there is one monitor in each cell, and the remaining monitors are clustered in one cell.

3.2.4 Extension to other utilities

In this chapter we have discussed considerations related to the selection of utility functions to represent preferences for both higher values of the process of interest Z, and good spatial coverage of the region. Many of the principles discussed may be extended to different choices of utility function, as the situation

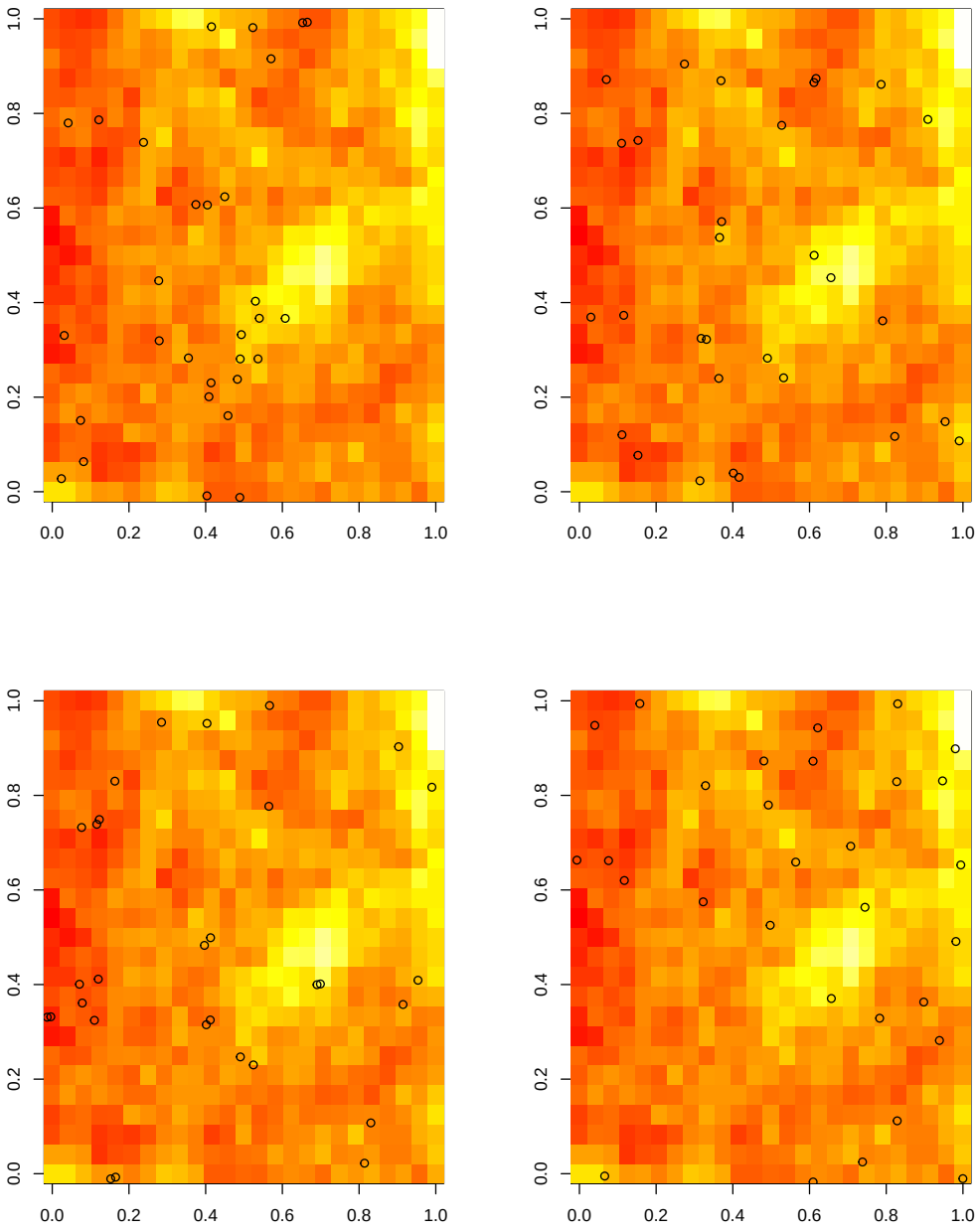


Figure 3.3: Four designs, the top two of which have the same mean-nearest-neighbour distance, but different mean-coverage distance. The bottom two show two designs that are equivalent in terms of their mean-coverage distance, but have different mean-nearest-neighbour distances.

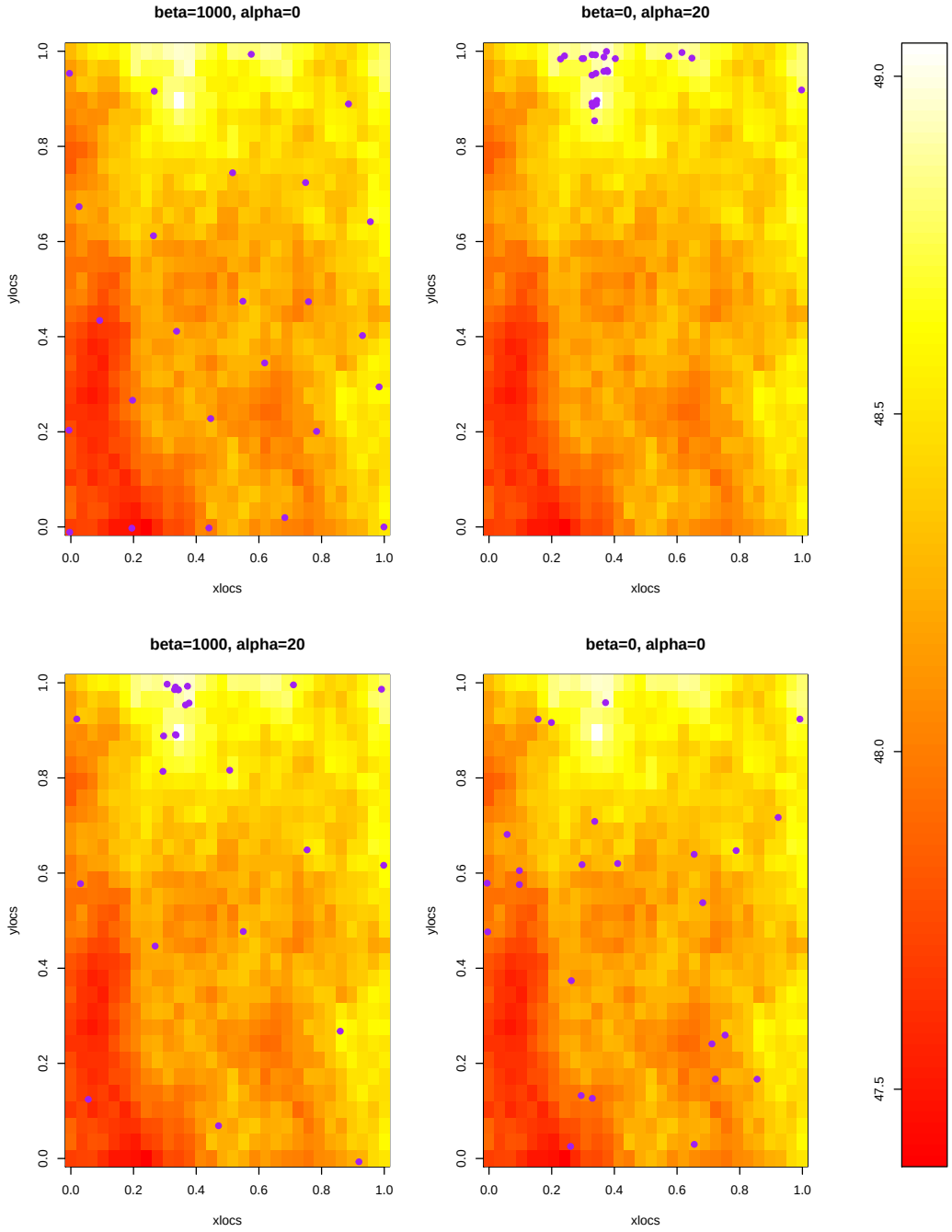


Figure 3.4: Generated designs from the mean distance utility with varying α and β.

necessitates. For example, a preference for space-filling may be balanced with a preference for having monitors close to certain pollutant point sources, or population centres, as in Gelfand et al. (2012). In such a situation, the distance from the pollutant source can also be included as a predictor for Z, meaning the utility function would take on a subtly different role, as the lower estimation further from the sampling locations (as would be informed by the knowledge that preferential sampling has gone on), is at least partially already dealt with by the knowledge that they are further from the point source of pollution. In this situation the advantage of accounting for preferential sampling would lie in the fact that the experimenter may have knowledge of other processes going on, which are not clear to the modeller.

Another avenue for exploration could be to include the relative costs associated with sampling from different areas, due to access issues, building regulations, appropriateness of terrain etc.. In this chapter we have considered balanced space-filling and preference utilities, to address the situation in which space-filling preferences mask otherwise strong preferences for higher values. We may encounter the situation in which similarly strong preferences for high values are masked by higher monitor placement costs in those areas.

Likewise, suppose we are interested in a quantity measured (possibly preferentially) only at specific points along a specific road, and we wish to predict the values along the whole stretch of road. Rather than using the space-filling functions described above, it may be more appropriate to define a one dimensional non-Euclidean 'along-road distance' space-filling function as part of the utility. In such a situation similar considerations about the use of mean, maximum and minimum distances must be made.

Further extensions could include having a utility function which is dependent on a previous sampling network. It may be costly to move a monitoring station, and thus an experimenter may assign a higher utility to designs that have monitoring sites in common with previous sampling networks. In such a situation, if different monitoring networks have been combined, or if new monitors have been introduced for different purposes it may be important to consider whether separate utilities should be assigned to different subsets of monitors. Such an idea is discussed further in Chapter 7.

Chapter 4

Estimating the normalising constant of the design distribution

We discussed in Chapter 1 the need for the ratio of normalising constants of distributions implied by utility functions and the difficulty in calculating them exactly for all but the smallest of cases. We shall expand on this further in Sections 5.1 and 5.2, when we update Z or α while fitting a joint model to the spatial and sampling process via an MCMC algorithm. We will require the ratio of the normalising constants for the design distributions implied by different values of Z and the strength of preference parameter(s) α. We will focus the discussion on the case where Z is updated and α is fixed. The same follows when updating α with Z fixed. We seek to replace the ratio $\frac{K(Z_1;\alpha)}{K(Z_2;\alpha)}$ with an approximation. In this chapter we shall consider a selection of methods for constructing such an approximation, along with the methods for sampling from utility-implied design distributions which are necessary for such approximations. We shall consider a special class of utility functions, the properties of which enable better normalising constant estimation. Finally we shall discuss a method of incorporating these normalising constant ratio estimates into a wider MCMC algorithm.

4.1 Generating design samples

The approximation methods which we shall employ require that we can draw samples of the design from a distribution proportional to the utility function. In this section we focus on the task of sampling designs from

$$P(D|Z,\alpha) = \frac{U(Z,D;\alpha)}{\sum_{D\in\mathbb{D}} U(Z,D;\alpha)}.$$

We associate two vectors with each design: a length N count vector D containing the number of samples taken at each site, as usual, and a length n index vector $\mathbf{q}$ containing the locations of the sample sites. i.e., for $N = 5$, $n = 3$, $\mathbf{q} = (1, 2, 3)$ would imply $D = (1, 1, 1, 0, 0)$.

4.1.1 Metropolis-Hastings point-swapping

One site at a time

The simplest design sampling method involves taking any arbitrary design and moving one site at a time, accepting or rejecting the change according to the Metropolis Hastings ratio. A scaling factor takes into account the probabilities associated with the different number of ways of moving from one design to another, due to repetitions of sites within the initial and new designs and the fact that we treat monitors as indistinguishable.

Algorithm:

1. Initialise with design $\mathbf{q}^0$.
2. At iteration i randomly select a site from $\mathbf{q}^{i-1}$ to be changed, with each of the n sites having uniform probability of selection, and replace it with another, to make $\mathbf{q}^*$.
3. Set $\mathbf{q}^i = \mathbf{q}^*$ with probability
$$\min\left\{1, \frac{U(D^*, Z; \alpha)}{U(D^i, Z; \alpha)} \frac{v_o}{v_n}\right\},$$
where v_o is the number of repeats of the site to be changed that there were in the original, current-state design, and v_n is the number of repeats of the site that was changed to in the new, proposed design. This fraction accounts for the transition probabilities. Otherwise set $\mathbf{q}^i = \mathbf{q}^{i-1}$.
4. Repeat from step 2.

This algorithm is very simple and therefore each step is very fast. However, as only one step is taken at a time it takes many steps to converge, with each draw likely to be very highly correlated with the previous one. This problem is worse for higher values of n as changing one site at a time is a proportionally smaller step size. In some cases the length of the burn-in period for this design sampling algorithm can be reduced in a wider MCMC algorithm in which these design samples are involved in the drawing of new values of Z, by setting $\mathbf{q}_0$ as the final design sample corresponding to the last value of Z. Naturally, in this case, a sufficient (albeit smaller) burn-in period must still be used, otherwise, for example, where the ratio of utility function normalising constants is to be calculated by importance sampling using samples related to the value of Z in the denominator (as in Section 4.4), we are in danger of over-inflating the ratio as the last value of Z will be the value corresponding to the numerator. We may encounter problems related to multimodality, in which case some low-utility designs must be traversed to reach a different mode. With only one site changing at a time, these are unlikely to be accepted.

Many sites at a time

We can implement a similar algorithm, but changing more than one site at a time, analogous to increasing the step size. The way in which our acceptance ratio is scaled for repeated sites, is more complicated due to the fact that the ordering in a design does not matter, and that there may be many ways to

move between two equivalent designs by making one step.

Theorem: Let $\mathbf{q}^o$ be the original (current state) design vector, $\mathbf{q}^n$ the proposed design vector, and D^o and D^n their corresponding count vectors. We define the change vector $\mathbf{c} = D^n - D^o$. The transition probability ratio may be given as

$$\frac{\mathrm{P}(D^o \to D^n)}{\mathrm{P}(D^n \to D^o)} = \prod_{i=1}^{N} \frac{d_i^n!}{d_i^o!} = \frac{\prod_{i:c_i \neq 0} d_i^n!}{\prod_{i:c_i \neq 0} d_i^o!}. \tag{4.1}$$

Proof: We begin by using the fact that the ratio of the probabilities of moving from one design to another is the ratio of the number of different ways in which we can move from the first to the second design in one step to the number of designs that are a single step away, including multiple counting for designs which we can move to in more than one way.

Denoting the number of sites to be changed at any one time t, we can pick $\binom{n}{t}$ different combinations of locations within $\mathbf{q}^o$ to be changed. There are then N^t options for what can be put in these t gaps. We remark that neither of these two quantities depends on what the actual sites included within the designs are, so it is the same for $D^n \to D^o$ and $D^o \to D^n$, meaning that the transition probability ratio simplifies to

$$\frac{\mathrm{P}(D^o \to D^n)}{\mathrm{P}(D^n \to D^o)} = \frac{\#\text{ways of moving } D^o \to D^n}{\#\text{ways of moving } D^n \to D^o}.$$

We look at the number of ways of moving from D^o to D^n. This can be factorised into the number of ways a particular selection of sites can be chosen for removal, multiplied by the number of different replacement combinations which would result in the new design.

First we look at the number of ways a particular indistinguishable selection could have been chosen to be swapped. The number of sites taken out of each kind is $|\mathbf{c}_i|$ for i *s.t.* $c_i < 0$. Meanwhile, the number of repeats of site i in the original design is d_i^o. The number of ways a particular selection can be chosen to be swapped is the product of the number of ways each of the sets of indistinguishable sites could have been chosen. For example, if three out of five of site i, and five out of seven of site j had been chosen for swapping, there would have been $\binom{5}{3}\binom{7}{5}$ ways in which this selection could have been made. This leads to

$$\#\text{Ways of selecting sites for removal} = \prod_{i:c_i<0} \binom{d_i^o}{|\mathbf{c}_i|}.$$

Having taken t sites out of the design there is now a gap of length t to be filled. We look at the sites which have been added in. Say there are l distinct sites. They will each have multiplicity $|\mathbf{c}_i|$ for $i : c_i > 0$ in the change vector. As the ordering in D^n does not matter, we need to count the different orderings of the sites as different ways of arriving at the same design. This is equivalent to counting

the number of permutations of the t elements:

$$\frac{t!}{\prod_{i:c_i>0}(|\mathbf{c}_i|!)}.$$

Thus, overall we have

$$\#\text{ways of moving } D^o \to D^n = \frac{t! \prod_{i:c_i<0} \binom{d_i^o}{|\mathbf{c}_i|}}{\prod_{j:c_j>0}(|\mathbf{c}_j|!)}.$$

And so by symmetry we have

$$\frac{\mathrm{P}(D^o \to D^n)}{\mathrm{P}(D^n \to D^o)} = \frac{\left(\prod_{i:c_i>0} \binom{d_i^n}{|\mathbf{c}_i|}\right)\left(\prod_{j:c_j>0}(|\mathbf{c}_j|!)\right)}{\left(\prod_{i:c_i<0} \binom{d_i^o}{|\mathbf{c}_i|}\right)\left(\prod_{j:c_j>0}(|\mathbf{c}_j|!)\right)}.$$

Cancelling gives

$$\frac{\mathrm{P}(D^o \to D^n)}{\mathrm{P}(D^n \to D^o)} = \frac{\left(\prod_{i:c_i>0} d_i^n!\right)\left(\prod_{i:c_i<0} d_i^n!\right)}{\left(\prod_{i:c_i>0} d_i^o!\right)\left(\prod_{i:c_i<0} d_i^o!\right)}.$$

Now using the fact that for $i : s.t.\ c_i = 0\ D_i^n = D_i^o$

$$\frac{\mathrm{P}(D^o \to D^n)}{\mathrm{P}(D^n \to D^o)} = \prod_{i=1}^{N} \frac{d_i^n!}{d_i^o!}$$

□

Algorithm: We now describe an algorithm using this theorem for drawing a chain of samples of the design, in which multiple sites may be switched at each step.

1. Initialise with starting design $\mathbf{q}^0$.
2. Select n_c locations to be swapped in $\mathbf{q}^i$.
3. Remove them from $\mathbf{q}^i$ and replace them with a random sample of the same size from $1, \ldots, N$ to construct $\mathbf{q}^*$.
4. Calculate a scaling factor of transition probabilities, as given in Theorem 4.1.
5. Set $\mathbf{q}^{i+1} = \mathbf{q}^*$ with probability

$$\mathrm{MH}_{\mathrm{prob}} = \min\left\{1, \frac{\mathrm{P}(D^* \to D^i)}{\mathrm{P}(D^i \to D^*)} \times \frac{U(D^*, Z; \alpha)}{U(D^i, Z; \alpha)}\right\},$$

   otherwise set $\mathbf{q}^{i+1} = \mathbf{q}^i$.
6. Repeat from step 2.

As with the 'one site swap at a time' algorithm, this is very fast, yet has the potential to converge faster, as the steps taken between designs is larger. However, if too many sites are changed at once then the acceptance rate will be lower. Naturally the optimal choice of stepsize is situationally dependent: we have found that for situations in which the utility describes very strong preferences, even when only one site is changed at a time, the acceptance rate is often very low. Otherwise, we have aimed (via trial-and-error tuning of the number of sites to be switched) for an acceptance rate of around $20\% - 40\%$.

4.1.2 Independent multinomial proposals

In this section we consider the use of samples from a related multinomial distribution to draw design samples from different, utility-implied distributions. We consider utility functions of the form

$$U(D, Z; \alpha) = \frac{n!}{\prod_{i=1}^{N} d_i!} \exp\left(\sum_{i=1}^{N} \alpha d_j z_j + f(D)\right).$$

The combination utility (3.8) is an example of such a function. To obtain a design sample at iteration t we can propose a design D^* from the multinomial distribution with probabilities proportional to $\exp(\alpha z_j)$ for each site j, and accept or reject them with the probability

$$\exp(f(D^*) - f(D^{t-1}))$$

The acceptance rate, (i.e. the effectiveness in proposing useful designs) will be heavily dependent on how close this ratio is to 1.

Algorithm:

1. Initialise with starting design D^0.
2. At iteration i calculate a vector of probabilities for the current value of Z, $\mathbf{p} = \{p_1, \ldots, p_N\}$, s.t. $p_i \propto \exp(\alpha z_i)$.
3. Propose a value D^* from the multinomial distribution with probabilities defined by the vector $\mathbf{p}$.
4. Set $D_i = D^*$ with probability

$$\min\left\{1, \exp(f(D^*) - f(D^{i-1}))\right\}.$$

5. Repeat from step 3.

4.1.3 Sampling designs with non-reversible Markov chains.

We investigate whether it might be more efficient to sample designs using a non-reversible Markov chain. Recall, a Markov chain is reversible with respect to density $\pi(x)$ in the space χ if the transition probabilities $K(y, x)$ are such that

$$\pi(x)K(x, y) = \pi(y)K(y, x),$$

for all $x, y \in \chi$. This is a sufficient but not necessary condition for $\pi(x)$ to be the stationary distribution. Andrieu and Livingstone (2019) and Diaconis et al. (2000) describe an algorithm (first in one dimension) which embeds the distribution $\pi(x)$ within another one: $\tilde{\pi}(z, x)$ with $z \in \{-1, 1\}$ on a space $\chi \times \{-1, 1\}$ for which, when $z = -1$ we propose steps in a negative direction, and when $z = 1$ we propose steps in a positive direction. Then, if the proposal is accepted, the sign of z is switched with probability θ, whereas if the proposal is rejected the sign of z is switched with probability $(1 - \theta)$, where $\theta \in (0, 1)$ is some parameter. The effect of this, when θ is small, is that we keep moving the chain in the same direction until we get a rejection. Proofs that this forms an irreducible, aperiodic Markov chain with stationary distribution $\tilde{\pi}(z, x) = \frac{\pi(x)}{2}$ (which is what we require, as we are only interested in the x component), can be found in Diaconis et al. (2000), who also show that for uniform $\pi(x)$ this algorithm exhibits favourable convergence properties when compared with random walk chains. The algorithm is generalised to more than one dimension, with a $+1$ or -1 direction index for each direction we can move in, and in the case of sampling designs, takes the following route: we define an *arc* between every pair of distinct cells, each with a corresponding direction. This direction information can be stored in an antisymmetric direction matrix E of zeros (only on the diagonal), $+1$s and -1s, where $E_{ij} = 1, E_{ji} = -1$ indicates that the arc between cells i and j goes from cell i to cell j. In this case we call i the root and j the tip of the arc. The algorithm, with weights $w_{1,2}, \ldots, w_{N-1,N}$, which may, for example, take the same values, or $w_{ij} \propto d_i + d_j$ etc., and switching probability parameter θ governing the probability of spontaneous reversal, proceeds as follows:

Algorithm: at iteration t

1. Select an arc (i, j) according to the probabilities w_{ij}.
2. Propose that a monitor be moved from the root of the arc i to the tip of the arc j, to form a proposal design D^* with $d_i^* = d_i^t - 1, d_j^* = d_j^t + 1$.
3. With probability $\alpha = \min\left(1, \frac{U(D^*,Z)}{U(D^t,Z)}\right)$, set $d_i^{t+1} = d_i^*$, $d_j^{t+1} = d_j^*$,otherwise let $d_i^{t+1} = d_i^t$, $d_j^{t+1} = d_j^t$. If the move proposed is impossible (i.e. because $d_i^t = 0$) then $\alpha = 0$. If the move is rejected $E_{ij} = -E_{ij}$ and $E_{ji} = -E_{ji}$. If the move is accepted the elements of E remain unchanged.
4. With probability θ, switch the direction of the arc: $E_{ij} = -E_{ij}$ and $E_{ji} = -E_{ji}$.

While Diaconis et al. (2000) show that such non reversible methods give favourable convergence properties when compared with random walk Metropolis-Hastings, with respect to the uniform distribution, this may not hold for other distributions. We investigate this in the following section.

4.1.4 Quality of samples

In this section we propose a process for comparing the quality of chains of design samples generated in different ways. The general idea is to find a minimum spacing k, within a chain, such that the cell counts in each design j are as different from those in design $j + k$ as they would be, had they come from different chains. This may be achieved by running two chains in parallel, calculating the mean sum of squares between the cell counts at each possible value for $k = 1, 2 \ldots$ (which we shall define as $C_{k,i} = \frac{1}{n_k} \sum_j (d_i^j - d_i^{j+k})^2$, for cell i, and where there are n_k available pairs from which to calculate this).

These may then be combined into a statistic $C_k = \sum_i C_{k,i}^2$, and the spacing k increased until this value is within some appropriate level of tolerance of the equivalent value, when calculated using pairs from opposite chains (i.e $C_i^* = \frac{1}{n_k}\sum_j (d_{i,1}^j - d_{i,2}^j)^2$ with the second subscript term indicating the chain from which the cell count is taken, with the same number of design samples, n_k, and $C^* = \sum_i C_i^{*2}$). This is equivalent to finding the spacing required to reduce the estimated between draw covariance to be as small as it might be if it were theoretically zero, as with independent samples. This is because we have

$$\frac{1}{2}\mathrm{E}((d_i^{j+k} - d_i^j)^2) = \frac{1}{2}(\mathrm{E}((d_i^{k+j})^2) + \mathrm{E}((d_i^j)^2) - 2\mathrm{Cov}(d_i^{j+k}, d_i^j) - 2\mathrm{E}(d_i^j)\mathrm{E}(d_i^{j+k})) \tag{4.2}$$

$$= \mathrm{Var}(d_i) - \mathrm{Cov}(d_i^{j+k}, d_i^j). \tag{4.3}$$

The natural question that arises is what this 'appropriate level of tolerance' should be. We propose repeating this process of finding the between-chain mean squared difference value to find a 'bootstrap confidence interval' for the value this statistic should take in the situation when the designs are independent of one another. This involves calculating C^* repeatedly, say u times: $C^{*1}, \ldots, C^{*u}$, for each of which the second chain has undergone a random reordering. We may then select, say $v = \lfloor 0.9u \rfloor$, and take the value of the smallest integer k such that $C_k < C_v^*$. In other words, the k-spaced designs from the same chain fulfil this criterion of being 'as different' as if they had been from different chains. We can then define an 'effective sample size' as the length of the chain, divided by this spacing k. An alternative method of estimating the effective sample size is to use the `multiESS` function in the `mcmcse` package which estimates the effective sample size of a multivariate chain, as described in Vats et al. (2019), from the covariance structure of the system of variables. In our specific case this means treating each of the N sites as a different variable. This becomes problematic as there are some sites which are never or rarely sampled, meaning that there is rarely enough information on how they vary with other variables and the Markov chain is not large enough for the asymptotics to work, thus causing the function to fail. While one option is to exclude these almost-never sampled sites, this may lead to ambiguous results: while it may be the case that these sites have very low probability in the true distribution, it may also be that the space has not been properly explored. If this occurs with too many sites, the function also fails.

We have predominantly used the 'bootstrap effective sample size' method for comparing design-sampling methods. It is also useful to carry out this process on preliminary runs of the wider MCMC algorithm, to get an idea of an appropriate sample size of design samples for each step, given some starting estimates for values of Z etc..

Example: comparison of sampling methods

We use the 'bootstrap effective sample size' procedure to compare samples drawn from the combination utility (3.8), with Z a Gaussian random field and $N = 15 \times 15 = 225$, $\theta = 5$ $\sigma = 4$, $\varphi = 1.5$. We compare samples generated using the one-site-switching Metropolis Hastings (M.H. 1 site), three-site-switching Metropolis Hastings, (M.H. 3 site), independent multinomial proposal (I.P.), and non-reversibe chain (N.R.) in methods for a variety of values of α, β and n. In order to make designs with different values of n more comparable in terms of strength of preference, we scale β by $\sqrt{n}$ and α by $\frac{1}{n}$, as the sum

Parameters	M.H.(1 site)	M.H.(3 sites)	N.R.	I.P.
$\alpha = 200/n$, $\beta = 0$	1.15	0.239	0.160	100
$\alpha = 200/n$, $\beta = 20\sqrt{n}$	0.629	0.398	0.249	9.09
$\alpha = 200/n$, $\beta = 100\sqrt{n}$	0.621	0.758	0.356	0
$\alpha = 200/n$, $\beta = 200\sqrt{n}$	0.735	0.847	0.177	0
$\alpha = 200/n$, $\beta = 400\sqrt{n}$	0.420	0.190	0.331	0
$\alpha = 100/n$, $\beta = 0$	0.633	1.052	0.617	100
$\alpha = 100/n$, $\beta = 10\sqrt{n}$	0.361	0.893	0.435	14.3
$\alpha = 100/n$, $\beta = 50\sqrt{n}$	0.685	0.870	0.690	2.94
$\alpha = 100/n$, $\beta = 100\sqrt{n}$	0.826	0.870	0.238	0.877
$\alpha = 100/n$, $\beta = 200\sqrt{n}$	0.658	0.439	0.457	0.289
$\alpha = 50/n$, $\beta = 0$	0.857	0.160	0.160	100
$\alpha = 50/n$, $\beta = 5\sqrt{n}$	0.833	1.04	0.248	91.0
$\alpha = 50/n$, $\beta = 25\sqrt{n}$	0.610	1.79	0.599	7.14
$\alpha = 50/n$, $\beta = 50\sqrt{n}$	0.592	0.971	0.376	14.28
$\alpha = 50/n$, $\beta = 100\sqrt{n}$	0.481	0.962	0.452	0.892

Table 4.1: Comparison of sampling methods, results for designs with 100 sampled locations. The values displayed are the effective sample sizes, as a percentage of the whole chain.

of sampled Z cells scales with n and closest distances by $\frac{1}{\sqrt{n}}$. We use (relative to n) three values of $\alpha : \frac{50}{n}, \frac{100}{n}, \frac{200}{n}$ and β values of $0, \frac{n^{\frac{3}{2}}\alpha}{10}, \frac{n^{\frac{3}{2}}\alpha}{2}, n^{\frac{3}{2}}\alpha, 2n^{\frac{3}{2}}\alpha$. This has the overall effect that for each of the values of α (which are roughly comparable over the differing values of n) we have values of β which mean the designs range from completely multinomial, to almost completely grid-like. The sets of experiments corresponding to different values of α then correspond to different strengths of preference for both high values and space-filling. We have specified, for this experiment, that for the bootstrap effective sample size procedure, the within sample covariance must be within the 90^{th} percentile of the between-sample covariance. It takes roughly the same time to get the same sized sample from each method. For each experiment 20000 samples were used with the first 2500 discarded. The results are shown in Table 4.1 and Table 4.2.

We first note that there is a higher acceptance rate for weaker preference. This is to be expected as the distribution is closer to a uniform distribution, in which every move would be accepted. Next, we can see that generating designs using independent multinomial proposals tends to be either very good or very bad, with a very high acceptance rate where the distribution is close to a multinomial distribution, and, to a lesser extent, where both preferences are very weak. For higher strengths of preference it is not a good method to use, as there is a zero acceptance rate unless the distribution is completely multinomial. In this example there is not any substantial difference demonstrated between the results for non reversible and random walk methods, suggesting that the potential gains from using a non reversible chain, described by Diaconis et al. (2000) do not carry over to this situation. Finally, we can see that, where we use random walk proposals and where we have a higher value of n, it is beneficial to switch more sites at once, shown by the better performance of the three-site switch Metropolis Hastings algorithm, compared with the one-site switch algorithm for $n = 100$, and the opposite being seen for $n = 20$. While there will, most likely, be situations for which each of these methods is most

Parameters	M.H. (1 site)	M.H.(3 sites)	N.R.	I.P.
$\alpha = 200/n$, $\beta = 0$	3.23	0.287	0.193	100
$\alpha = 200/n$, $\beta = 20\sqrt{n}$	1.09	0.242	0.153	0
$\alpha = 200/n$, $\beta = 100\sqrt{n}$	0.508	0.0504	0.249	0
$\alpha = 200/n$, $\beta = 200\sqrt{n}$	0.160	0.102	0.120	0
$\alpha = 200/n$, $\beta = 400\sqrt{n}$	0.300	0.408	0.279	0
$\alpha = 100/n$, $\beta = 0$	1.59	1.00	0.457	100
$\alpha = 100/n$, $\beta = 10\sqrt{n}$	1.47	0.671	0.459	10.0
$\alpha = 100/n$, $\beta = 50\sqrt{n}$	0.629	0.538	0.137	0
$\alpha = 100/n$, $\beta = 100\sqrt{n}$	0.515	0.188	0.0696	0
$\alpha = 100/n$, $\beta = 200\sqrt{n}$	0.424	0.249	0.280	0
$\alpha = 50/n$, $\beta = 0$	2.632	4.00	0.355	100
$\alpha = 50/n$, $\beta = 5\sqrt{n}$	1.39	1.37	0.325	11.1
$\alpha = 50/n$, $\beta = 25\sqrt{n}$	0.917	1.16	0.389	0.877
$\alpha = 50/n$, $\beta = 50\sqrt{n}$	0.901	0.296	0.255	0.408
$\alpha = 50/n$, $\beta = 100\sqrt{n}$	0.549	0.339	0.329	0.150

Table 4.2: Comparison of sampling methods, results for designs with 20 sampled locations. The values displayed are the effective sample sizes, as a percentage of the whole chain.

appropriate, henceforth we shall tend to use random walk Metropolis Hastings with a small number of site-switchings at each move. It is worth noting that the estimated effective sample sizes, which are relatively small, are both a function of n and N (as when these are higher it takes longer for each possible site to have a chance of being included in the design, and for each site within in the design to have the chance to be removed), and that they are likely to be overly conservative estimates. This conservatism comes from the fact that, when there is not a sufficiently large burn-in period, samples not from the target distribution will be included in the opposite-chain covariance estimators, which will thus be underestimated, as there will naturally be less similarity between the cell counts for those and the cell counts in samples from the opposite chain.

4.1.5 Conclusions

In this section we have considered and compared methods for generating samples of designs from distributions proportional to utility functions. Clearly the optimal choice of sampling method is highly situationally dependent: for utilities very close to multinomial distributions there is a clear advantage of close to i.i.d. sampling, which is lost when the distribution is further away. Likewise, with point swapping algorithms, the optimality of number of sites to change depends very heavily on n, N and how peaked the distribution is. As we can see in Table 4.1 there is a danger of a very low acceptance rate when switching too many sites at once.

4.2 Estimating the normalising constant ratio

4.2.1 Importance sampling

Importance sampling using MCMC samples

The first method of normalising constant estimation we consider uses design samples generated, via MCMC from one of the utility-implied distributions in question. Say we are interested in finding the ratio

$$\frac{K(Z_1,\alpha)}{K(Z_2,\alpha)} = \frac{\sum_{D\in\mathbb{D}} U(D,Z_1;\alpha)}{\sum_{D\in\mathbb{D}} U(D,Z_2;\alpha)}, \tag{4.4}$$

and have access to samples $D_1,\dots D_M$ from an ergodic Markov chain with target distribution $P(D|Z_2,\alpha)$. Then we have

$$\begin{aligned}
\frac{1}{M}\sum_{i=1}^{M}\frac{U(D_i,Z_1,\alpha)}{U(D_i,Z_2,\alpha)} &\xrightarrow{n\to\infty} \mathrm{E}_{D\sim P(D|Z_2,\alpha)}\left(\frac{U(D_i,Z_1,\alpha)}{U(D_i,Z_2,\alpha)}\right)\\
&= \sum_{D\in\mathbb{D}}\frac{U(D_i,Z_1,\alpha)}{U(D_i,Z_2,\alpha)}\frac{U(D_i,Z_2,\alpha)}{K(Z_2,\alpha)}\\
&= \frac{\sum_{D\in\mathbb{D}} U(D_i,Z_1,\alpha)}{K(Z_2,\alpha)}\\
&= \frac{K(Z_1,\alpha)}{K(Z_2,\alpha)}.
\end{aligned}$$

Thus we have a consistent estimator for the ratio $\frac{K(Z_1,\alpha)}{K(Z_2,\alpha)}$. The use of such an estimator is a key part of the Monte Carlo Metropolis Hastings (MCMH) method of Liang and Jin (2013) which we shall discuss later.

Importance sampling using related multinomial samples

A further option is the estimation of the normalising constant $K(Z)$ using samples from a related multinomial distribution. Due to the fact that the above distribution integrates to 1, with multinomially sampled $d_1\dots d_N$ from $P(D|Z,\alpha)$ a multinomial distribution as in (2.1) we have

$$\frac{1}{N}\sum_{i=1}^{N}\frac{U(D_i,Z_1;\alpha)}{P(D_i|Z_1;\alpha)} \xrightarrow{N\to\infty} K(Z_1;\alpha).$$

We expect this method to perform well when the utility function has a significant multinomial component, e.g. in the case of a combination space-filling utility, where there is only very weak preference for space-filling. Likewise, this method may present an advantage when N and n are large enough for it to be difficult to produce independent samples of designs from the utility-implied distributions. However, the comparative performance of this method may be different when calculating the ratio of constants, dependent on a variety of factors:

1. How close to one another the Z vectors are: the closer they are then the better the performance of the methods using samples from the utility distributions will be. If they are very far apart

then each of the utility functions may be closer to their corresponding multinomial distributions than to each other, leading to comparatively better estimation using the multinomial sampling method.

2. Closeness of the utility function to a multinomial distribution: it is worth noting that this does not necessarily need to be the multinomial distribution implied by the utility function with the same strength of preference parameter α: in the case of a combination utility, a lower value of α^* may be more appropriate. However, the selection of an optimal α^* value is not straightforward, and itself may be computationally expensive.

3. The size of n and N will have a large effect on the quality of the Metropolis-Hastings samples.

4.2.2 Reverse logistic regression

If there is little agreement between Z_1 and Z_2, the importance sampling methods may lead to estimators with a high variance, due to the fact that there may be little overlap of the support of high probability regions of the two distributions. In this case each $\frac{U(D,Z_1)}{U(D,Z_2)}$ may be either very large or very small, leading in turn to a high variance in the estimator. A technique described by Geyer (1994) known as reverse logistic regression (RLR) may be of use in this case. This involves drawing samples from multiple distributions $f_1, \ldots, f_n$, the normalising constant ratios of which we require. We draw m_j samples from each of $j = 1 \ldots n$ distributions before "pretending to forget" which distribution each of the samples x_{ij}, $i = 1, \ldots, m_j$ came from. The probability that x came from distribution j is

$$p_j(x) = \frac{m_j f_j(x)}{\sum_{k=1}^{n} m_k f_k(x)}.$$

We define $f_j(x) = \frac{g_j(x)}{K_j}$, $\mathbf{r} = (r_1, \ldots, r_m)$, and $r_k = \frac{K_1}{K_k}$ and write

$$p_j(x, \mathbf{r}) = \frac{m_j r_j g_j(x)}{\sum_{k=1}^{n} m_k r_k g_k(x)}.$$

We then maximise the pseudo-loglikelihood that each sample came from its own distribution with respect to $\mathbf{r}$:

$$l(\mathbf{r}) = \sum_{j=1}^{n} \sum_{i=1}^{m_j} \log(p_j(x_{ij}, \mathbf{r})),$$

which will give us an estimate for $\mathbf{r}$. As with importance sampling, we require there to be at least some level of overlap between the two distributions in question. If this is not the case, then the p_j terms will be very close to one, as the probability of an observation having come from the other distribution will be negligible, and the pseudolikelihood will be a constant. An advantage of this method lies in our ability, if necessary, to introduce more distributions with greater overlap between the two of interest, such as those implied by previous iterations.

4.2.3 Example: estimating a ratio $\frac{K_1}{K_2}$

The relative performance of the methods which use multinomial samples depend on how 'multinomial-like' these designs are. We illustrate this by way of an experiment. We construct two Gaussian random fields Z_1 and Z_2, realised at points over a 5×5 regular grid on the unit square. Z_1 has mean zero, correlation parameter $\log(\varphi) = 2.5$ and variance parameter $\sigma^2 = 0.3$ and exponential correlation function. Z_2 is Z_1 with added Gaussian noise with mean zero, variance 0.1. We generate 20 such Z_1 and Z_2 realisations, and allowing $n = 7$ monitors and utility (3.8) with $\alpha = 3$, calculate the normalising constant ratio $\frac{K(Z_2)}{K(Z_1)}$ exactly. Then, with 1000 design samples, we estimate the normalising constant ratios using each of the three methods (RLR, importance sampling with both Metropolis Hastings and multinomial samples) and calculate the mean squared error for each method. The distributions used for RLR were those which were proportional to the utility function (3.8) for values Z_1 and Z_2 for Z. We repeat this for a range of values of β. The results are shown in Figure 4.1. Clearly as β increases, it becomes less advisable to calculate the normalising constant ratio using multinomial samples, while the methods using RLR and importance sampling with the Metropolis Hastings generated samples maintain better (and very similar) levels of accuracy. This suggests that the method of using multinomial samples should only be used for cases in which there is, firstly, a difficulty in producing the samples from the utility-implied distributions, and secondly, when the distribution is very close to a multinomial distribution, i.e. there are only very weak preferences for things such as space-filling.

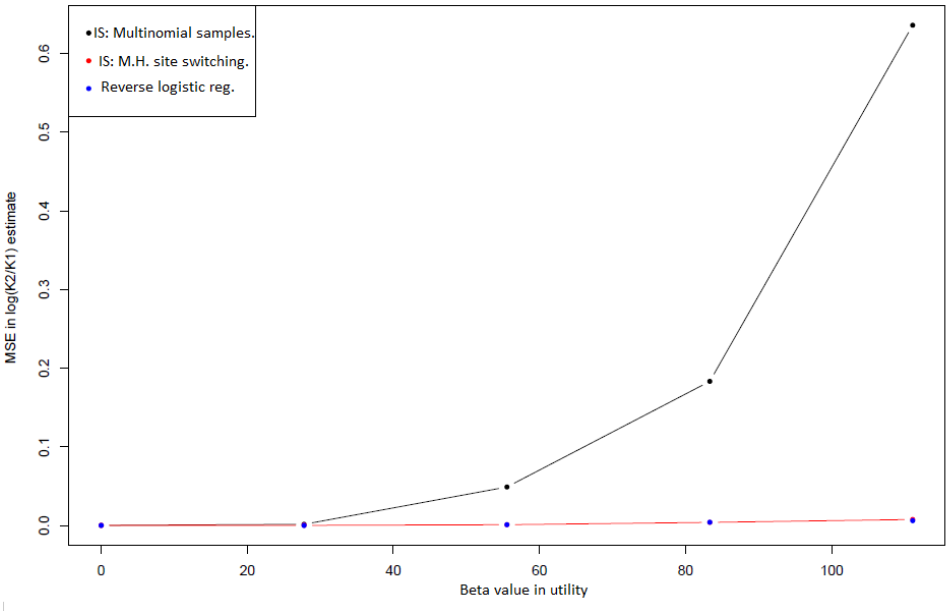


Figure 4.1: Comparison of mean squared errors of the log of the normalising constant ratios for three methods of estimation, with $N = 25$, $n = 7$ and utility function (3.8).

4.3 Permutation invariant utility functions

In this section we define and investigate a special class of utility functions, the properties of which allow us to employ a method which can lead to a reduction of variance in the estimator of the ratio

of their related normalising constants. Naturally, the modeller should choose a utility function which best mirrors the real-life process in question, yet if this utility is a member of this class, such efficiency-improving techniques can be applied.

More specifically, these are functions, the sums (over the whole design space) of which are unchanged when the elements of the Z argument are arbitrarily reordered. By reordering we may bring the two Z values and their corresponding utility functions closer together, and in doing so we may reduce the variance of the estimator of their normalising constant ratios. Provided the variance of estimator may be reduced in this way, we may, in turn, reduce the number of samples with which it is constructed without detrimental effect on the inferences made on Z. For high dimensional Z and D this could prove very useful: the generation of designs is by far the most time-consuming part of any MCMC algorithm by which we predict Z (see Section 5.1).

We proceed by defining these functions, before exploring what kind of functions fit this definition; we state some conditions under which we may combine these functions, give some practical examples of some of these functions, and demonstrate their usage and usefulness by way of several examples.

We define a *permutation invariant utility function* $U(D, Z)$ to be a utility function such that

$$\sum_{D \in \mathbb{D}} U(D, \phi Z) = \sum_{D \in \mathbb{D}} U(D, Z), \tag{4.5}$$

where ϕZ is some permutation of $Z \in \mathbb{R}^N$ s.t. $(\phi Z)_i = z_{\phi(i)}$ and $D \in \mathbb{D}$. We omit any parameters α from our notation in this section, as they will not be used. Recall, our goal is to find an approximation of the normalising constant ratio

$$\frac{K(Z_1)}{K(Z_2)} = \frac{\sum_D U(D, Z_1)}{\sum_D U(D, Z_2)}.$$

The methods of achieving this encounter difficulties when Z_1 and Z_2 are very different and the densities proportional to their corresponding utility functions have very little overlap. By reordering Z_1 and Z_2 to be as similar as possible, while not altering the normalising constant, we can lessen this problem by bringing $U(D, Z_1)$ and $U(D, Z_2)$ closer together. In order to do this we require the functions $K(Z)$ to be invariant under permutation of Z.

4.3.1 What kind of functions are permutation invariant?

We seek utility functions U which satisfy (4.5). We can partition $\mathbb{D}$ into r disjoint classes $\mathbb{D}^1, \ldots, \mathbb{D}^r$ such that the elements of each $D \in \mathbb{D}^i$ are the same, albeit in a different order. For example, where $N = 3$, $n = 3$ one class would be

$$\mathbb{D}^i = \{(0,1,2), (0,2,1), (1,2,0), (1,0,2), (2,1,0), (2,0,1)\}.$$

As $\mathbb{D}$ may be divided into such classes, it is sufficient for (4.5) to seek functions for which

$$\sum_{\mathbb{D}^i} U(D, Z) = \sum_{\mathbb{D}^i} U(D, \phi Z) \quad \forall i \in \{1, \ldots r\}.$$

Say we consider, without loss of generality,

$$\sum_{\mathbb{D}^1} U(D, Z),$$

we want every term in this sum to appear as a term in the sum

$$\sum_{\mathbb{D}^1} U(D, \phi Z),$$

and vice versa, i.e. for every design $D^j \in \mathbb{D}^1$ we require that there be a $D^k \in \mathbb{D}^1$ s.t.

$$U(D^j, Z) = U(D^k, \phi Z).$$

As each design in a class is a permutation of the others, this condition is satisfied if for each ϕ there exists a permutation ψ on the elements of each $D \in \mathbb{D}^1$ such that

$$U(\psi D, \phi Z) = U(D, Z).$$

We now look specifically at functions of the form

$$U(D, Z) = l(\sum_{k=1}^{N} g_k(D) z_k), \tag{4.6}$$

and investigate what properties of such functions would ensure permutation invariance. We require that for every ψ there is a ϕ such that

$$\sum_{k=1}^{N} g_k(\psi D) z_{\phi(k)} = \sum_{k=1}^{N} g_k(D) z_k.$$

Trivially, we have

$$\sum_{k=1}^{N} g_{\phi(k)}(D) z_{\phi(k)} = \sum_{k=1}^{N} g_k(D) z_k,$$

where $\{g_{\phi(1)}, \ldots, g_{\phi(N)}\}$ is the reshuffling of $\{g_1, \ldots, g_N\}$ under permutation ϕ and so we require that for every ϕ there be a ψ such that

$$\sum_{k=1}^{N} g_k(\psi D) z_{\phi(k)} = \sum_{k=1}^{N} g_{\phi(k)}(D) z_{\phi(k)}.$$

As there are no restrictions on Z, this satisfied by requiring that for each $k \in \{1, \ldots, N\}$

$$g_k(\psi D) = g_{\phi(k)}(D). \tag{4.7}$$

We look at three separate cases in relation to these functions $g_1(D), \ldots, g_N(D)$.

Case 1 : All the functions $g_1(D), \ldots, g_N(D)$ are symmetric in D.

Proposition: Given a set of symmetric-in-D functions $g_1(D), \ldots g_N(D)$ there exists a permutation ψ on D for every permutation ϕ on Z such that (4.7) holds if and only if $g_i(D) = g_j(D) \quad \forall j, i$.

Proof:

Say we have a permutation ψ for every ϕ such that (4.7) holds. Say $\phi(k) = j$ (without loss of generality), then we would have

$$g_{\phi(k)}(D) = g_j(D) = g_k(\psi D) = g_k(D),$$

by the symmetry: thus, these functions $g_i(D)$ must all be the same, as k and j may be chosen arbitrarily. Likewise if all the functions are the same then

$$g_i(D) = g_j(D) \quad \forall i, j,$$

So we have

$$g_k(\psi D) = g_k(D) = g_{\phi(k)}(D),$$

which is property (4.7). □

Case 2: At least one, but not all of the functions $g_1(D) \ldots g_N(D)$ are symmetric in D.

Proposition: Given a set of functions $g_1(D), \ldots, g_N(D)$ for which this property holds, there does not exist a permutation ψ on D corresponding to each permutation ϕ on Z such that (4.7) holds.
Proof: Say

$$g_k(D) = g_k(\psi D),$$

but there was a $j \neq k$ such that

$$g_j(D) \neq g_j(\psi D),$$

for some permutation ψ, i.e. g_j is nonsymmetric in D. This means

$$g_j(D) \neq g_k(D),$$

because they must be different functions for one to be symmetric and one to be nonsymmetric. Say ϕ is such that $\phi(k) = j$. Then we require the existence of a specific permutation ψ^* on D such that

$$g_k(\psi^* D) = g_{\phi(k)}(D) = g_j(D),$$

but then we would also have

$$g_k(D) = g_j(D),$$

which is a contradiction. □

Case 3: None of the functions $g_1(D), \ldots, g_N(D)$ is symmetric in $D = (d_1, \ldots d_N)$.

Proposition: A set of functions $\{g_1(D), \ldots, g_N(D)\}$ which are not symmetric has a permutation ψ on D for every permutation ϕ on Z such that (4.7) holds if and only if it is closed under the action of any permutation ψ on D: i.e. $\forall\psi\{g_1(D), \ldots, g_N(D)\} = \{g_1(\psi D), \ldots, g_N(\psi D)\}$
Proof:

1. Closure of the set of functions $\{g_1(D), \ldots, g_N(D)\}$ under permutation ψ on $D \implies$ (4.7) .

   We have that the set of functions $\{g_1(D), \ldots, g_N(D)\}$ is closed under permutation ψ, and we want to show that this means that every ϕ has a corresponding ψ such that

   $$g_{\phi(i)}(D) = g_i(\psi D),$$

   for every i.

   We begin by showing that two distinct permutations cannot correspond to the same ϕ. Say we have two permutations ψ_1 and ψ_2 with $\psi_1 \neq \psi_2$, that both satisfy (4.7).

   So for every $i \in \{1, \ldots, N\}$ we have

   $$g_{\phi(i)}(D) = g_i(\psi_1 D) = g_i(\psi_2 D),$$

   so, in terms of the elements $d_1, \ldots d_N$ of D, we have

   $$g_i(d_{\psi_1(1)}, \ldots, d_{\psi_1(N)}) = g_i(d_{\psi_2(1)}, \ldots, d_{\psi_2(N)}),$$

   for all D. Say g_i is a function of only $d_1, \ldots, d_k$, a subset of $\{d_1, \ldots, d_N\}$.Then we have

   $$d_{\psi_1(j)} = d_{\psi_2(j)},$$

   for all $j \in \{1, \ldots, k\}$. This must hold for all values of i (i.e. every one of the g-functions) and every element of D must be included in at least one g-function, as they are closed under permutation of d and none of the g functions is zero as none of them is symmetric in D. So in that case ψ_1 and ψ_2 must act upon all elements of D in the same way, which is a contradiction.

   Now we have that each of the $N!$ permutations ψ on D leads to a different permutation ϕ on the g-functions - so every ϕ must have a corresponding ψ, as required.

2. (4.7) $\implies$ closure under any permutation ψ on D.

   We have that for every permutation ϕ on the set of g-functions there is an 'equivalent' permutation ψ on D such that

   $$g_{\phi(j)}(D) = g_j(\psi D),$$

for all D. There are $N!$ possible distinct permutations ϕ on g and clearly, each must have a different ψ, of which there are also $N!$, so they must each simply reorder the g functions, so the set of g functions must be closed under permutation of D.

In summary, we have that for functions of the form (4.6) we require either that the functions $g_1(D), \ldots g_N(D)$ are all equal, or that they are all nonsymmetric and closed under permutation on D.

4.3.2 Combining permutation invariant utility functions

We may wish to combine utility functions. Clearly, the sum of permutation invariant utility functions is also a permutation invariant utility function. If we have

$$\sum_{D \in \mathbb{D}} U_i(D, \phi Z) = \sum_{D \in \mathbb{D}} U_i(D, Z),$$

for $i \in \mathbb{N}$ then we have

$$\sum_{D \in \mathbb{D}} \sum_{i} U_i(D, \phi Z) = \sum_{i} \sum_{D \in \mathbb{D}} U_i(D, \phi Z) = \sum_{i} \sum_{D \in \mathbb{D}} U_i(D, Z) = \sum_{D \in \mathbb{D}} \sum_{i} U_i(D, Z).$$

Proposition: The product of two permutation invariant utility functions is itself permutation invariant if for both of them there is the same mapping between permutations ϕ on Z and permutations ψ on D such that:

$$U(\phi Z, \psi D) = U(Z, D).$$

Proof. Say we have two utility functions U_1 and U_2 such that for every permutation ϕ there is the same permutation ψ on D such that

$$U_1(\psi D, \phi Z) = U_1(D, Z) \text{ and } U_2(\psi D, \phi Z) = U_2(D, Z),$$

then clearly we have a ψ s.t.

$$U_1(\psi D, \phi Z) U_2(\psi D, \phi Z) = U_1(D, Z) U_2(D, Z).$$

Thus $U_1 U_2$ is itself permutation invariant. □

Examples:

1. Multinomial sampling.

   The utility function which is equivalent to multinomial sampling is

   $$U(D, Z) = \frac{1}{d_1! \ldots d_N!} \exp(\alpha \sum_{i=1}^{N} d_i z_i),$$

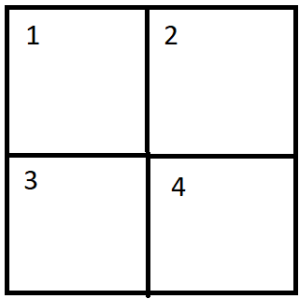
   which may be phrased as two utility functions: $U_1(D, Z) = \frac{1}{d_1! \ldots d_N!}$, and $U_2(D, Z) = \exp(\alpha \sum_{i=1}^{N} d_i z_i)$. The second of these is of the form (4.6) with $g_i(D) = \alpha d_i$, and so the set of functions $\{g_1(D), \ldots, g_N(D)\}$

is clearly closed under any permutation, so this utility U_2 is permutation invariant. As we also have, by symmetry, $U_1(D, Z) = U_1(\psi D, \phi Z)$ for any choices of ϕ and ψ we can apply the above proposition about the product of permutation invariant utilities (Section 4.3.2) to conclude that the utility $U(D, Z)$ is permutation invariant.

2. $$U(D, Z; M, \beta) = \sum_{i=1}^{N} z_i d_i + \beta \sum_{k=1}^{N} \sum_{j=1}^{N} d_i d_j M_{ij},$$

where M is the Euclidean distance matrix, is a permutation invariant utility function as each $g_i(D) = d_i$, the set of which is invariant under permutation of D, and the second term does not involve Z at all, and so is uanffected by permutation on it.

3. Say we had the following 2×2 mesh:



and the utility function was such that the usefulness of a monitor in a square was lessened by the presence of one in a neighbouring square, for example

$$\begin{aligned} U(D, Z) = (d_1 - \frac{1}{2}(d_2 + d_3))z_1 + (d_2 - \frac{1}{2}(d_1 + d_3))z_2 \\ + (d_3 - \frac{1}{2}(d_1 + d_4))z_3 + (d_4 - \frac{1}{2}(d_2 + d_3))z_4, \end{aligned} \tag{4.8}$$

which has $g_1(D) = (d_1 - \frac{1}{2}(d_2 + d_3))$ etc. This is not permutation invariant, as, for example, the set of g-functions is not invariant under the swapping of d_1 and d_2 : there is no $g_i = d_3 - \frac{1}{2}(d_4 + d_2)$.

4.3.3 Permutation invariant utility functions and estimating the normalising constant ratio

As earlier described, this consideration of permutation invariant utility functions is motivated by the fact that if there is a great deal of discrepancy between Z_1 and Z_2 (which imply the utility functions U_1 and U_2 respectively will be very different) then, in the case of Reverse Logistic Regression, the probability that a sampled design came from a distribution other than its own may be virtually zero,

leading to unidentifiability of the normalising constant ratios. Likewise, there will be high variance in importance sampling estimators. The advantage of permutation invariant utility functions lies in our ability to manipulate Z_1 and Z_2 to bring their corresponding utility functions closer together.

Example: variance of Reverse logistic regression estimators of the normalising constant ratios.

We have two vectors Z_1 and Z_2 of 25 i.i.d. normal observations with mean 0 and standard deviation 1, as displayed in matrix form in Figure 4.2 along with sorted vectors of the same values. Sorting the vectors makes them more element-wise similar.

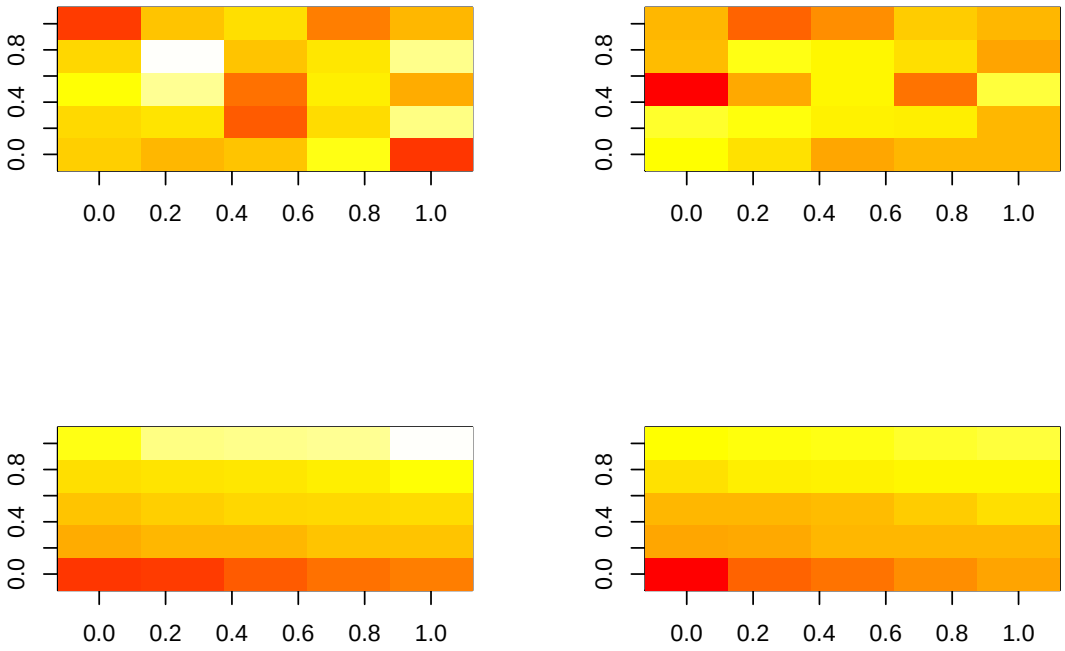


Figure 4.2: Sorted and unsorted Gaussian values Z_1 and Z_2

We conducted an experiment to demonstrate the effect of calculating the normalising constant ratio via Reverse Logistic Regression using the sorted Zs. We use a utility function proportional to a multinomial utility:

$$U(Z, D, \alpha) = \frac{1}{\prod_{j=1}^{N} d_i!} \exp\left(\alpha \sum_{i=1}^{N} d_i z_i\right),$$

with $\alpha = 0.5$. For this function we are able to calculate the normalising constant ratio exactly, for comparison. We define there to be $n = 20$ monitors. In this case we have $\log(K_2/K_1) = -2.88$.

We estimate the log of the normalising constant ratio 1000 times, using 4000 samples from each utility implied distribution (generated via one-step Metropolis Hastings, with a 1000 sample burn-in).

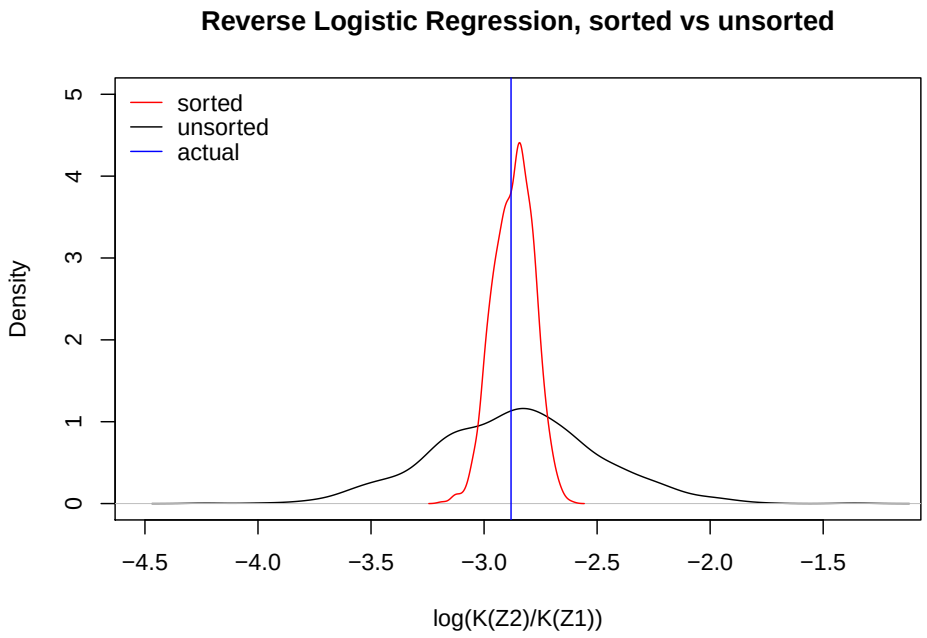


Figure 4.3: Density plot of log ratios of normalising constant estimators, estimated via Reverse Logistic Regression for sorted and unsorted Z_1 and Z_2.

We do this for both the sorted and unsorted methods. The densities of the resulting estimators are displayed in Figure 4.3, with the true log ratio -2.88 shown by the vertical line. The original (unsorted) method yielded a mean estimator of -2.86, standard deviation 0.361, while the new (sorted) method yielded a mean estimator of -2.87, with standard deviation of 0.0871, which is clearly much smaller. We expect similar variance reductions when using importance sampling with Metropolis-Hastings generated samples. Our estimator in this case is:

$$\frac{1}{m}\sum_{i=1}^{m}\frac{U(D_i,Z_2)}{U(D_i,Z_1)},$$

where $D_1 \dots D_m$ are samples generated from the density in the denominator

$$P(D|Z_1) \propto U(D,Z_1).$$

While reordering Z_1 and Z_2 should have no effect on $K(Z_2)/K(Z_1)$ when we have a permutation invariant utility function, and thus no effect on the expectation of this estimator, we have (assuming independence of samples)

$$\mathrm{Var}\left(\frac{1}{m}\sum_{i=1}^{m}\frac{U(D_i,Z_2)}{U(D_i,Z_1)}\right) = \frac{1}{m}\mathrm{Var}\left(\frac{U(D,Z_2)}{U(D,Z_1)}\right)$$

and while the expectation will still be the same, a higher discrepancy between Z_1 and Z_2 will lead to the

average being calculated from a mixture of very high and very low values, leading to a higher variance in the estimator.

Example: variance of importance sampling estimators of the normalising constant ratios

We repeat the above experiment, with the same parameters and the same log ratio $\log(K2/K1) = -2.88$, but this time using the importance sampling estimator for the log ratio of normalising constants. We use 4000 designs samples from the multinomial utility implied by Z_1. The density plots are shown in Figure 4.4 with the true ratio represented by the vertical red line. The sorted method gave a mean estimator value of -2.89 with standard deviation 0.158, while the original, unsorted method gave a mean estimator of -3.85 with standard deviation 1.09. Clearly this is an effective technique for increasing the

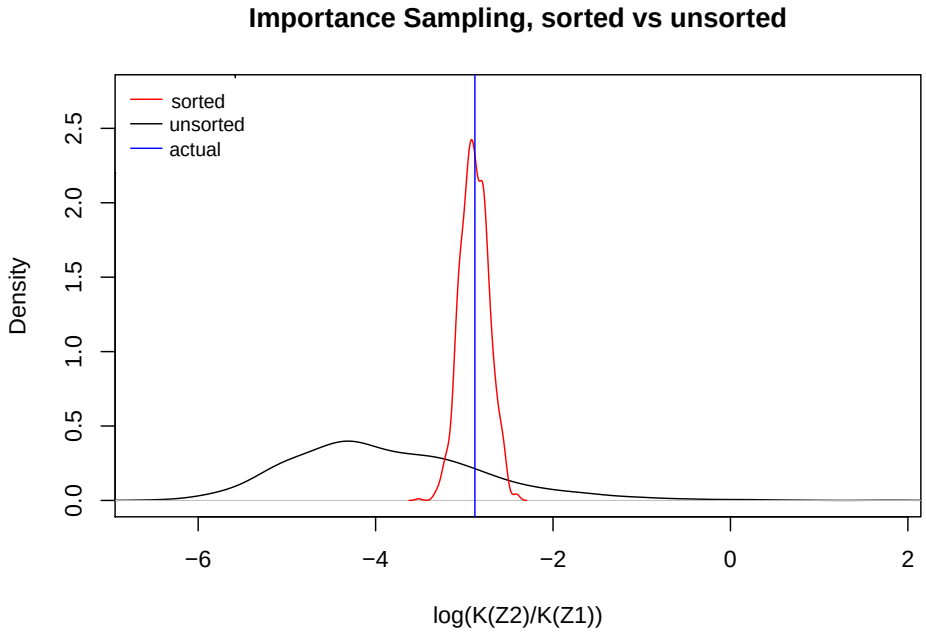


Figure 4.4: Densities of the log ratios of normalising constant estimators, estimated using importance sampling with Metropolis Hastings generated samples, both with sorted and unsorted Z_1 and Z_2.

accuracy of our estimators for the ratios of normalising constants: Under an assumption of independence of design samples, this would correspond to each term in the estimator sums having variance of 100 and 4760 for the 'sorted' and 'unsorted' methods respectively, which would imply that if we were using the 'sorted' method we could use a sample size roughly 48 times smaller. In fact, we repeat the same experiment with 8 samples (rather than 4000) after the same burn-in of 1000. In this situation the variance of the sorted method estimator with 8 samples is still lower (0.935) than the unsorted method estimator with 4000 samples. Even with the relatively large burn-in of 1000 samples, this represents a very substantial reduction in computation time. This may prove very useful in the wider MCMC algorithm in which inference on Z is performed, although the factor by which time is saved will depend on other computationally intensive parts of the algorithm, such as correlation matrix inversions.

4.3.4 Permutation semi-invariant utility functions

Not all useful utility functions share this property, such as many of those depending on distances between cells, via a Euclidean a distance matrix M. Nonetheless, using similar principles, we may define a *permutation semi-invariant* utility function as one for which we have:

$$\sum_{D\in\mathbb{D}} U(D, Z; M) = \sum_{D\in\mathbb{D}} U(D, \phi Z; \phi M)$$

where ϕ is a permutation. When ϕ acts on a matrix it acts on both row and column indices. Say we have, associated with ϕ, a permutation matrix Φ. We then have

$$\phi Z = \Phi Z,$$

$$\phi M = \Phi M \Phi^T.$$

Similarly, using the same logic as above, a function may be permutation semi-invariant if for every permutation ϕ on Z and M we have a permutation ψ on D s.t.

$$U(\psi D, \phi Z, \phi M) = U(D, Z, M).$$

The combination space-filling and preferential sampling utility (3.8) is an example of such a function.

Example: Combination space-filling and preferential sampling utility

We consider the combination utility function (3.8). Clearly the ratio $\frac{n!}{\prod_i d_i!}$ is unchanged under any permutation of D. Next, we note that the permutation ψ of D which must be applied to keep high value preference term $\alpha \sum_{i=1}^{N} z_i d_i$ unchanged when we apply permutation ϕ to the elements of Z is also ϕ. Consequently, if the space-filling term

$$g(D) = g_{MC}(D) + g_{NN}(D),$$

with $g_{MC}(D)$ and $g_{NN}(D)$ as in (3.7) and (3.6) respectively, is unchanged when the elements of both the design D and the distance matrix M are acted upon by the same permutation ϕ, then we have permutation semi invariance. It is straightforward to show that this is the case, as the sets of values over which the means and minima are to be taken are simply relabelled in the same way as the elements of M.

Finding an optimal permutation

Unlike with permutation invariant utility functions, there is no straightforward method such as putting Z in size order which will bring the individual utilities for each sampled design closer together, as bringing the Z values closer may push the actual utilities further apart by changing the matrix M. We can think of this as an assignment problem: for every position in the new Z and M ordering we must assign an index from the old ordering. However, as the 'rewards' for each index-position combination depend on the positions of the other indices, it cannot be phrased as a linear programming problem.

One, albeit inefficient, option is repeatedly switching pairs of indices, calculating the variance, and accepting a switch if the variance is reduced. To improve the efficiency, we can choose to increase the number of swaps proposed for the most important sites within the sample we already have i.e. the is for which d_i appears most frequently. To do this we restrict one of the sites to be swapped to be from the set of sites which were sampled most often. While such a method might seem promising, numerical experimentation shows that the gains in terms of variance reduction were rarely worth the extra computational cost.

4.3.5 Combining sorted-method and unsorted-method estimators

An alternative approach to attempting to find the optimal permutation involves sampling designs from both $U(D, Z_2; M, \alpha)$ and $U(D, \Phi Z_2; \Phi M, \alpha)$ where Φ is the permutation which reorders Z_2 to be in the same size order as Z_1. With $D_1 \dots D_L \sim U(D, Z_2; M, \alpha)$ and $D_1^* \dots D_L^* \sim U(D, \Phi Z_2; \Phi M, \alpha)$ we have

$$\frac{1}{2L}\sum_{i=1}^{L}\frac{U(D_i, Z_1; M, \alpha)}{U(D_i, Z_2; M, \alpha)} + \frac{1}{2L}\sum_{i=1}^{L}\frac{U(D_i^*, Z_1; M, \alpha)}{U(D_i^*, \Phi Z_2; \Phi M, \alpha)} \xrightarrow{L\to\infty} \frac{K(Z_1;\alpha)}{K(Z_2;\alpha)}.$$

Simulation results

We ran simulations to test whether sampling from both distributions (with sorted and unsorted Zs and Ms) reduces the variance of the estimator of the ratio $\log\left(\frac{K(Z_1;\alpha)}{K(Z_2;\alpha)}\right)$. We generate two Gaussian random fields with mean zero, correlation parameter $\log(\varphi) = 2$, variance $\sigma^2 = 5$ and exponential correlation function, with realisations on a 12×12 grid on the unit square. We set there to be $n = 8$ monitors and use the combination utility (3.8) with $\alpha = 4.5$ and $\beta = 150$. Naturally, for this utility function we cannot calculate the exact $\log\left(\frac{K_1}{K_2}\right)$ value, but for the original, unsorted method and the proposed, half-sorted method we estimate this value 200 times, using a total of 5000 samples for each, generated via Metropolis Hastings, proposing to move one monitoring station at each iteration. For the unsorted method the mean value for the estimator was -6.31, with variance 146, whereas the values for the half-sorted method were -3.188 and 68.98 respectively. The density plots are shown in Figure 4.5. While this method does not give as dramatic reductions in variance as is possible for permutation invariant utility functions, this example demonstrates that it still may be a worthwhile step, as it is relatively computationally inexpensive, compared with generating more design samples: we repeat the above experiment with iteratively fewer design samples and found that, for the data in question, one has to use fewer than 30 samples on average when using the half-sorted method for the variance of those estimators to have comparable variance to those produced by the 'unsorted' method with 5000 samples. To put this into perspective, for this particular example, the generation of 1000 samples takes approximately 480 milliseconds, while sorting the Z vector takes approximately 70 microseconds on the same machine, which is a comparatively negligible computation time. Obviously these computational gains vary according to the specific situation: we have found that methods that involve sorting work better when the ratio of α to β is higher. Where these computation time gains are of great importance it may be worthwhile setting estimated thresholds, dependent on α and β to determine what proportions of the sample (if any) should come from the 'sorted' or 'unsorted' methods. However, developing a

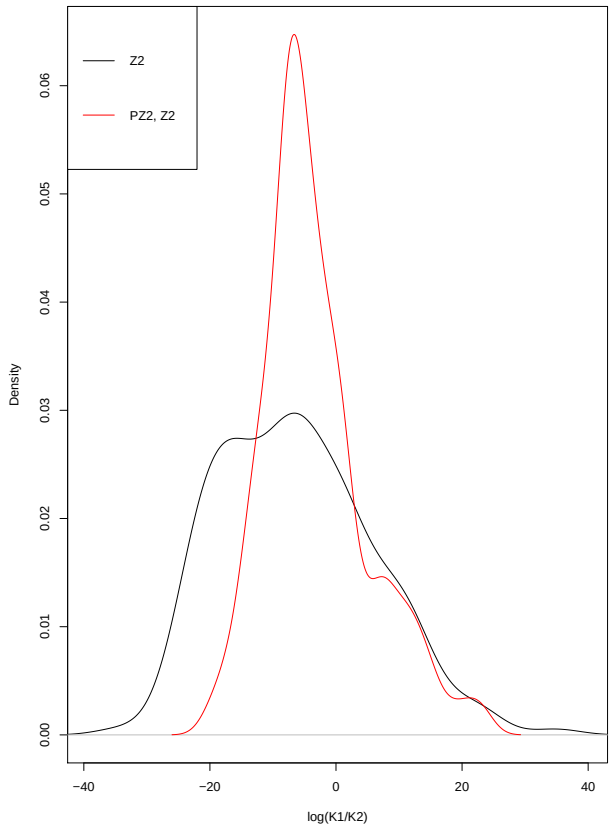


Figure 4.5: Density plots for the log K-ratio estimators, when the 'unsorted' and 'half-sorted' methods are used, for generated data as in 4.3.5 , for a permutation semi-invariant utility function.

method of determining what these thresholds should be requires more investigation.

4.4 Monte Carlo Metropolis Hastings

In this section we consider a method, proposed by Liang and Jin (2013), known as Monte Carlo Metropolis Hastings (MCMH) for drawing samples from a posterior distribution (which in our case would be for Z or α) which makes use of an importance sampling estimator. We present it first in a general context. The model under consideration is given by

$$\theta \sim \pi(\theta)$$

$$x|\theta \sim f(x|\theta)$$

where

$$f(x|\theta) = \frac{g(x|\theta)}{K(\theta)},$$

where $K(\theta) = \int g(x|\theta)dx$ is an intractable normalising constant of $f(x|\theta)$. This leads to posterior distribution $\pi(\theta|x)$ with

$$\pi(\theta|x) = \frac{g(x|\theta)\pi(\theta)}{K(\theta)}.$$

The objective is to sample θ from $\pi(\theta|x)$. We require the ability to calculate

$$\frac{g(x|\theta_1)K(\theta_2)}{g(x|\theta_2)K(\theta_1)},$$

for two different values θ_1 and θ_2. However, this is not possible where we cannot calculate $K(\theta)$. The idea is to replace

$$\frac{K(\theta_1)}{K(\theta_2)}$$

for values θ_1 and θ_2 of θ, with a consistent estimator. If we have access to samples $x_1, \ldots, x_n \sim f(x, |\theta_2)$, say from an ergodic Markov chain, then we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\frac{g(x_i|\theta_1)}{g(x_i|\theta_2)} &\xrightarrow{n\to\infty} \mathrm{E}_{x\sim f(x|\theta_2)}\left(\frac{g(x_i|\theta_1)}{g(x_i|\theta_2)}\right) \\
&= \int \frac{g(x|\theta_1)}{g(x|\theta_2)}\frac{g(x|\theta_2)}{K(\theta_2)}dx \\
&= \frac{\int g(x|\theta_1)dx}{K(\theta_2)} \\
&= \frac{K(\theta_1)}{K(\theta_2)}.
\end{aligned}$$

Thus we have a consistent estimator for the ratio $\frac{K(\theta_1)}{K(\theta_2)}$. This leads to the following algorithm for sampling from the posterior distribution $\pi(\theta|x)$ of θ:

Algorithm:

1. Initialise with value θ^0.
2. At step t
   - Propose value $\theta^* \sim q(\theta|\theta^{t-1})$.
   - Generate $x_1, \ldots, x_n \sim f(x|\theta^{t-1})$.
   - Estimate $\widehat{\frac{K(\theta^*)}{K(\theta^{t-1})}} = \frac{1}{n}\sum_{i=1}^{n}\frac{g(x_i,\theta^*)}{g(x_i,\theta^{t-1})}$.
   - Set $\theta^t = \theta^*$ with probability

$$\tilde{\alpha}_n = \min\left(1, \frac{g(x,\theta^*)\pi(\theta^*)q(\theta^{t-1}|\theta^*)}{g(x,\theta^{t-1})\pi(\theta^{t-1})q(\theta^*|\theta^{t-1})}\widehat{\frac{K(\theta^{t-1})}{K(\theta^*)}}\right)$$

     otherwise set $\theta^t = \theta^{t-1}$.
3. Repeat from step 2.

Assuming that we could calculate $\frac{K(\theta)}{K(\theta^*)}$ exactly, the transition kernel for the MCMC algorithm, when transitioning from θ to θ^*, $\alpha(\theta, \theta^*)$ is given by

$$q(\theta^*|\theta)\alpha(\theta,\theta^*) + \left(1 - \int_\Theta q(\theta^*|\theta)\alpha(\theta,\theta^*)d\theta^*\right) 1(\theta^* = \theta), \tag{4.9}$$

where $\alpha(\theta,\theta^*) = \min\left\{1, \frac{K(\theta)}{K(\theta^*)}\frac{g(\theta^*,x)\pi(\theta^*)}{g(\theta,x)\pi(\theta)}\frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}\right\}$and Θ is the space of all θs. We assume that this has stationary distribution $\pi(\theta|x) \propto \frac{1}{K(\theta)}g(x,\theta)\pi(\theta)$. In the above proposed MCMH algorithm the transition kernel from $(\theta, \mathbf{x})$ to $(\theta^*, \mathbf{x}^*)$, with generated auxiliary data $\mathbf{x} = \{x_1, \ldots, x_n\}$, and auxiliary data $\mathbf{x}^* = \{x_1^*, \ldots, x_n^*\}$ relating to θ^*, is instead given by

$$\begin{aligned} P_n(\theta^*, \mathbf{x}^*|\theta, \mathbf{x}) &= f(\mathbf{x}^*|\theta^*)q(\theta^*|\theta^*)\tilde{\alpha}_n(\theta,\theta^*,\mathbf{x}) \\ &\quad + \left(1 - \int_\Theta q(\theta^*|\theta)\tilde{\alpha}_n(\theta,\theta^*,\mathbf{x})d\theta^*\right) 1(\theta^* = \theta)1(\mathbf{x} = \mathbf{x}^*), \end{aligned} \tag{4.10}$$

where $\tilde{\alpha}_n(\theta,\theta^*,\mathbf{x}) = \min\left\{1, \widehat{\frac{K(\theta)}{K(\theta^*)}}\frac{g(\theta^*,x)\pi(\theta^*)}{g(\theta,x)\pi(\theta)}\frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}\right\}$ and $\widehat{\frac{K(\theta)}{K(\theta^*)}}$ has been estimated using n samples. Multiplying the transition kernel (4.10) by $f(\mathbf{x}|\theta)$ and integrating with respect to both $\mathbf{x}$ and $\mathbf{x}^*$, the marginal transition kernel for θ is

$$\begin{aligned} P_n(\theta^*|\theta) &= q(\theta^*|\theta)\int \tilde{\alpha}_n(\theta,\theta^*,\mathbf{x})f(\mathbf{x}|\theta)d\mathbf{x} \\ &\quad + 1(\theta^* = \theta)\int\left(1 - \int_\Theta q(\theta^*|\theta)\tilde{\alpha}_n(\theta,\theta^*,\mathbf{x})d\theta\right) f(\mathbf{x}|\theta)d\mathbf{x}. \end{aligned}$$

It can be seen by the consistency of the estimator $\dfrac{\widehat{K(\theta^*)}}{K(\theta)}$ that

$$\int \tilde{\alpha}_n(\theta,\theta^*,\mathbf{x})f(\mathbf{x}|\theta)d\mathbf{x} \to \int \alpha_n(\theta,\theta^*)f(\mathbf{x}|\theta)d\mathbf{x} = \alpha_n(\theta,\theta^*),$$

as $n \to \infty$ and, consequently, $P_n(\theta^*|\theta)$ converges to the transition kernel (4.9). Liang and Jin (2013) prove rigorously convergence of the marginal chain for θ under the assumption that we are able to estimate $\dfrac{K(\theta)}{K(\theta^*)}$ consistently, and that we do not take arbitrarily large steps between θ and θ^*. That is, denoting the marginal distribution for θ, starting at an arbitrary value θ_0, after k iterations by $P_n^k(\theta|\theta_0)$ and the target distribution by $\pi(\theta|\mathbf{x})$, then

$$\|P_n^k(\theta|\theta_0) - \pi(\theta|\mathbf{x})\| \to 0 \text{ as } k, n \to \infty,$$

where $\|\cdot\|$ denotes the total variation norm.

We can apply this to our specific situation. Say we have two values Z_1 and Z_2, and their corresponding distributions:

$$P(D|Z_i,\alpha) = \frac{U(D,Z_i,\alpha)}{K(Z_i,\alpha)}.$$

Say we have access to samples $D^1, \ldots, D^n$ from $P(D|Z_1, \alpha) = \frac{U(D,Z_1;\alpha)}{K(Z_1;\alpha)}$. Then we have

$$\frac{1}{n}\sum_{i=1}^{n}\frac{U(D^i, Z_2, \alpha)}{U(D^i, Z_1, \alpha)} \xrightarrow{n\to\infty} \frac{K(Z_2, \alpha)}{K(Z_1, \alpha)},$$

which we may use in our calculation of the acceptance ratio for Z. We shall make use of this in the following chapters as we apply our methods to various data sets.

4.4.1 Sample sizes for normalising constant estimation

Generating samples of designs for normalising constant estimation is (along with inverting large matrices for sampling from posterior distributions of the covariance parameters) generally the most computationally intensive part of a wider MCMC algorithm used for making inferences about Z. As such, there is a trade-off (in terms of computational time) between the number of design samples generated at each step of the wider algorithm, and the number of overall steps taken.

Liang and Jin (2013) describe how, while the approximate transition kernel converges to the true transition kernel as the number of auxiliary samples m increases, (as in 4.4), this value m need not be very large: in the many examples they describe, a value of m between 20 and 50 gives very good results. Obviously this will be highly situationally dependent, and we have tended to use more auxiliary samples than this, due to the fact that there is often high correlation between design samples, especially for situations in which the utility functions represent strong preferences (as can be seen in Tables 4.1 and 4.2). While at the very least, when performing inference on Z, trace plots for the utilities of the generated designs $U(D^i, Z_1, \alpha)$, $i \in 1, \ldots, m$ should be examined, while tuning the algorithm, to ensure that the chain of design samples has sufficiently converged, it is not clear how to determine an optimal balance of the two sample sizes. Further investigations in this area could consider a scenario in which one has a total computational budget B, and wishes to find the optimal sample sizes L (for the outer MCMC), and M for the normalising constant estimation. Here, if the computational cost for each of the L MCMC samples and design samples were c_1 and c_2 respectively, we would have

$$c_1 L + c_2 LM = B,$$

and would choose L and M to minimise some measure of prediction discrepancy, such as the maximum prediction variance in the elements of Z, subject to this constraint. If the normalising constant were known, these variances would be approximately inversely proportional to L, but we do not have a precise formula for the sample variance in the case in which the normalising constants are estimated with some level of error. Instead, we propose an empirical method for evaluating the relationship of the prediction variance for different values of L and M based on short runs and use that to extrapolate to higher values of L and M.

This would involve estimating the criterion to be minimised (such as the prediction variance) by means of, for example, a non-overlapping batch means technique (Flegal and Jones, 2010) on these short initial runs, in which values $L^{'}$ and $M^{'}$, (analogous to L and M), were varied, subject to some smaller computational cost $B^{'}$. We could then choose the ratio of these two values $\frac{L^{'}}{M^{'}}$ which minimises

this, and use that as the ratio of $\frac{L}{M}$. From here we can find values of L and M, subject to the original constraint of total computational burden. It would be advisable to make sure these short runs made good exploration of the space of the strength of preference parameters, as these have a large effect on the convergence of the design-sampling chain. It may be possible to establish relationships analytically between the overall prediction variance and the variance of the normalising constant estimators, as a topic for future research.

Chapter 5

Application of the whole model to data

We now bring together the discussed ideas by describing a hierarchical model by which we can model jointly the spatial process Z and the sampling design D. We shall detail the process of fitting such a model with the combination utility (3.8), before applying this method to both simulated and real data sets in order to evaluate the effects of its use on the prediction of Z and other parameters. In particular, we shall compare, via simulation study, the relative performance (in terms of accurate recovery of Z) of the combination utility, with that of a multinomial utility (approximate LGCP, as in Diggle et al. (2010)). We shall then show the difference made in a real-world context by accounting for preferential sampling in this way via application to a data set of soil ammonia concentrations over a sheep-field in Scotland.

5.1 A hierarchical model for the spatial and design processes

The model we shall focus on assumes the spatial surface to be a mean θ Gaussian random field with exponential covariance structure with variance σ^2 and correlation parameter φ. Measurements $Y = \{y_i : i = 1, \ldots, n\}$ are taken at n locations, according to a design drawn from the utility-implied distribution, with parameters α controlling the strength of preference for high or low values, or other considerations. The elements of Y are assumed to be noisy measurements of $X = \{x_i : i = 1, \ldots, n\}$ of the Gaussian random field at the irregular sampling locations $\tilde{D}$, with mean zero i.i.d. Gaussian noise with variance τ^2:

$$y_i = X_i + \epsilon_i, \quad i = 1, \ldots, n$$

$$\epsilon_i \sim N(0, \tau^2), \quad \text{i.i.d.},$$

where X_i is the value of X at the location of the i^{th} point in the design. We wish to predict Z, the values of the Gaussian random field at a set of regular discretisation focus points, with which are also associated the cell counts D of the sampling design. The dependency structure of the random variables is displayed in Figure 5.1 and their joint distribution factorises as follows:

$$P(Z, X, D, Y, \alpha, \sigma^2, \varphi, \theta, \tau^2) = P(Z, X|\sigma^2, \varphi, \theta)P(D|Z, \alpha)P(Y|X, \tau^2)P(\alpha)P(\varphi)P(\sigma^2)P(\tau^2)P(\theta).$$

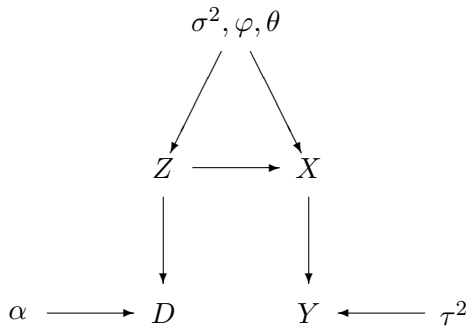


Figure 5.1: DAG displaying the dependence structure in our proposed model.

The full list of distributions which we will assume is as follows:

$$\alpha \sim N(0, h)$$

$$\varphi \sim IG(a, b)$$

$$\sigma^2 \sim IG(c, d)$$

$$\tau^2 \sim IG(f, g)$$

$$\theta \sim N(k, l)$$

where $a, b, c, d, f, g, h, k, l$ are hyperparameters.

$$(Z, X)|\varphi, \sigma^2 \sim N_{n+N}(\theta 1_{n+N}, C)$$

$$C_{ij} = \sigma^2 \exp(-|s_i - s_j|/\varphi), \quad i, j = 1, \ldots, n + N$$

where s_i and s_j are co-ordinates corresponding to the ith and jth elements of the vector (Z, X).

$$D|Z, \alpha \sim \frac{U(D, Z; \alpha)}{K(Z; \alpha)}.$$

Herein lies our assumption that the choice of sampling locations depends on the underlying field $\tilde{Z}$ only via its values Z at the focus points of each cell in the discretisation. In practice these focus points will be the cell centres. For the measurements Y we have

$$Y|X, \tau^2 \sim N_n(X, \tau^2 I).$$

Here we assume that, given the underlying values X of $\tilde{Z}$ at the chosen locations, the observations have i.i.d. Gaussian noise. As discussed in previous chapters, when fitting the model via Markov chain Monte Carlo a key difficulty is presented by the need for ratios of the often incalculable normalising constant $K(Z; \alpha)$, requiring the summation over all $\binom{N+n-1}{n}$ possible designs. We will address this problem using the Monte Carlo Metropolis Hastings method of Liang and Jin (2013). These

intractable normalising constants, in addition to presenting a challenge with respect to the acceptance ratios of Z and the parameters of the utility function, present a challenge in terms of formulating effective proposals for the often high-dimensional Z. We shall address this by the implementation of an approximate Metropolis-adjusted Langevin algorithm. In order to fit the hierarchical model we shall employ the following MCMC algorithm.

Basic MCMC algorithm

1. Initialise with starting values of X^1, Z^1, $\sigma^{2,1}$, φ^1, $\tau^{2,1}$, θ^1 α^1.

2. At step t

   - Draw X^t from

   $$P(X|Z^{t-1}, \sigma^{2,t-1}, \varphi^{t-1}, \theta^{t-1}, \alpha^{t-1}, \tau^{2,t-1}, D, Y) \propto P(X|\sigma^{2,t-1}, \varphi^{t-1}, Z^{t-1}, \theta^{t-1}) \times P(Y|X, \tau^{2,t-1}),$$

   which is a multivariate normal distribution by the symmetry of Y and X in $P(Y|X, \tau^2)$.

   - Draw Z^t from

   $$P(Z|X^t, \sigma^{2,t-1}, \varphi^{t-1}, \theta^{t-1}, \alpha^{t-1}, \tau^{2,t-1}, D, Y) \propto P(Z|X^t, \sigma^{2,t-1}, \varphi^{t-1}, \theta^{t-1}) P(D|Z, \alpha^{t-1}) \propto P(Z|X^t, \sigma^{2,t-1}, \varphi^{t-1}, \theta^{t-1}) \frac{U(D, Z; \alpha^{t-1})}{K(Z; \alpha^{t-1})}.$$

   (a) Propose Z^* from distribution $q(Z^*|Z^{t-1})$.

   (b) Set $Z^t = Z^*$ with probability

   $$\min\left\{1, \frac{P(Z^*|X, \sigma^{2,t-1}, \varphi^{t-1}, \theta^{t-1}) U(D, Z^*; \alpha^{t-1}) K(Z^{t-1}; \alpha^{t-1}) q(Z^{t-1}|Z^*)}{P(Z^{t-1}|X, \sigma^{2,t-1}, \varphi^{t-1}, \theta^{t-1}) U(D, Z^{t-1}; \alpha^{t-1}) K(Z^*; \alpha^{t-1}) q(Z^*|Z^{t-1})}\right\}, \quad (5.1)$$

   Otherwise set $Z^t = Z^{t-1}$.

   - Draw $\sigma^{2,t}$ from $P(\sigma^2|X^t, Z^t, \varphi^{t-1}, \alpha^{t-1}, \tau^{2,t-1}, \theta^{t-1}, D, Y) \propto P(X^t, Z^t|\sigma^2, \varphi^{t-1}, \theta^{t-1}) P(\sigma^2)$. If we assume inverse gamma priors for σ^2 this will simply be a draw from another inverse gamma distribution, as it is a conjugate prior for a normal distribution.

   - Draw φ^t from $P(\varphi|X^t, Z^t, \sigma^{2,t}, \alpha^{t-1}, \tau^{2,t-1}, \theta^{t-1}, D, Y) \propto P(X^t, Z^t|\sigma^{2,t}, \varphi, \theta^{t-1}) P(\varphi)$:

   (a) Propose φ^* from $q(\varphi^*|\varphi^{t-1})$

   (b) Set $\varphi^t = \varphi^*$ with probability

   $$\min\left\{1, \frac{P(X^t, Z^t|\sigma^{2,t}, \varphi^*, \theta^{t-1}) P(\varphi^*) q(\varphi^{t-1}|\varphi^*)}{P(X^t, Z^t|\sigma^{2,t}, \varphi^{t-1}, \theta^{t-1}) P(\varphi^{t-1}) q(\varphi^*|\varphi^{t-1})}\right\}.$$

   Otherwise set $\varphi^t = \varphi^{t-1}$

- Draw θ^t from $P(\theta|X^t, Z^t, \sigma^{2,t}, \alpha^{t-1}, \tau^{2,t-1}, \varphi^t, D, Y) \propto P(X^t, Z^t|\sigma^{2,t}, \theta, \varphi^t)P(\theta)$, which is another multivariate normal distribution.
- Draw $\tau^{2,t}$ from $P(\tau^2|X^t, Z^t, \sigma^{2,t}, \varphi^t, \alpha^{t-1}, \theta^t, D, Y) \propto P(Y|X^t, \tau^2)P(\tau^2)$. If we assume inverse gamma priors for τ^2 then this (as with σ^2) will be a draw from an inverse gamma distribution.
- Draw α^t from

$$\begin{aligned} P(\alpha|X^t, Z^t, \sigma^{2,t}, \varphi^t, \theta^t, \tau^{2,t}, D, Y) &\propto P(D|Z^t, \alpha)P(\alpha) \\ &= \frac{U(D, Z^t; \alpha)}{K(Z^t, \alpha)}P(\alpha). \end{aligned} \tag{5.2}$$

  (a) Propose α^* from $q(\alpha^*|\alpha^{t-1})$

  (b) Set $\alpha^t = \alpha^*$ with probability

$$\min\left\{1, \frac{U(D, Z^t, \alpha^*)K(Z^t, \alpha^{t-1})P(\alpha^*)q(\alpha^{t-1}|\alpha^*)}{U(D, Z^t, \alpha^{t-1})K(Z^t, \alpha^*)P(\alpha^{t-1})q(\alpha^*|\alpha^{t-1})}\right\}.$$

  Otherwise set $\alpha^t = \alpha^{t-1}$.

3. Repeat from Step 2.

The ratio $\frac{K(Z_1,\alpha_1)}{K(Z_2,\alpha_2)}$ appearing in the acceptance probability for Z and α is approximated using the methods described in Chapter 4. When the utility has more than one parameter, as in the combination utility function (3.8), each parameter is updated separately.

5.1.1 Sampling Z

Sampling the length N vector Z using random walk Metropolis Hastings is often inefficient for large values of N. This is because small changes in the vector can lead to large changes in the likelihood.

An alternative strategy might be the proposal of values from a multivariate normal distribution conditional on the the current values of X. These proposals would then be accepted according to the ratio of the current and proposed utilities and their normalising constants. While these proposals give good predictions of the higher valued, highly sampled regions, realisations of Z with significantly lower values in the unsampled areas will be proposed only with extremely low probability, thus the preferential sampling will not actually be accounted for.

We have found that a more successful approach to sampling Z uses a Langevin proposal distribution as part of the *Metropolis-adjusted Langevin algorithm (MALA)* (Roberts and Rosenthal, 1998), so called because Langevin dynamics are used to find the proposed value at each stage. This involves evaluating the gradient of the target distribution in order to travel in the direction of a region of higher probability density. While totally random walk Metropolis Hastings struggles to reach and sample from regions of high probability density for high dimensional random variables, MALA, by making proposals in the direction of the density gradient, directly 'aims for' regions of high probability density. There is much literature (such as that of Roberts and Tweedie (1996)) on how MALA algorithms have favourable convergence properties compared with random walk Metropolis Hastings algorithms, especially in high

dimensional settings. For example Roberts and Rosenthal (1998) show that, for certain n dimensional target distributions, a MALA algorithm can converge to its target distribution in $O(n^{\frac{1}{3}})$ steps where random walk Metropolis Hastings would require $O(n)$ steps. Likewise, they show that an optimal acceptance rate for a MALA algorithm is 0.574, as opposed to 0.234 for random walk Metropolis Hastings (i.e. less time need be taken up with the rejections associated with a step size big enough to explore the space adequately, as the steps are generally in a more helpful direction). Additionally, Pillai et al. (2012) show that this improved convergence also may be applied to n dimensional target distributions which are not necessarily the product of n independent one-dimensional distributions. In our case, as Z is of dimension N, these properties make MALA an attractive option to explore. Christensen et al. (2000) and Christensen et al. (2006) have demonstrated the use of these algorithms in a spatial modelling context. Our exploratory investigations of this method with data simulated from known Gaussian random fields showed that the Markov chains in question arrived at random fields reasonably close to the known truth fields much more quickly than when random-walk Metropolis Hastings.

The algorithm involves, given current state Z_t of a Markov chain, a proposal

$$Z^* = Z_t + \sigma_n \xi_{t+1} + \frac{\sigma_n^2}{2} \bigtriangledown \log(\pi_n(Z_t)),$$

where $\pi_n(Z)$ is the target density, ξ_{t+1} a value from a standard normal distribution, and σ_n^2 a tuning parameter. The acceptance probability for a Langevin proposal is

$$\min\left(\frac{\pi_n(Z^*)q(Z^*, Z_t)}{\pi_n(Z_t)q(Z_t, Z^*)}, 1\right),$$

where

$$q(x, y) = \frac{1}{(2\pi\sigma_n^2)^{n/2}} \exp(-\frac{1}{2\sigma_n^2}||y - x - \frac{\sigma_n^2}{2} \bigtriangledown \log(\pi(x))||_2^2).$$

In our specific case where our target density is the posterior distribution of Z:

$$P(Z|X, D, \alpha, \sigma, \varphi, \tau^2, \theta, Y) \propto P(Z|X, \sigma^2, \varphi, \theta)P(D|Z; \alpha),$$

we have

$$\bigtriangledown \log(P(Z|X, D, \alpha, \sigma, \varphi, \tau^2, \theta, Y)) = \bigtriangledown \log(P(Z|X, \sigma^2, \varphi, \theta)) + \bigtriangledown \log(P(D|Z; \alpha)).$$

Fairly straightforwardly

$$\begin{aligned}\bigtriangledown \log(P(Z|X, \sigma^2, \varphi)) = &-(\Sigma_{ZZ} - \Sigma_{XZ}^T \Sigma_{XX}^{-1} \Sigma_{XZ}^T)^{-1} Z \\ &+ (\Sigma_{ZZ} - \Sigma_{XZ}^T \Sigma_{XX}^{-1} \Sigma_{XZ}^T)^{-1} (\theta + \Sigma_{ZX} \Sigma_{XX}^{-1} (X - \theta)),\end{aligned}$$

where Σ_{ZZ}, Σ_{XX} and Σ_{XZ} are the covariance matrices for (Z, Z), (X, X) and (X, Z) respectively.

Things are slightly more complicated for the second term :

$$\bigtriangledown \log(P(D|Z;\alpha)) = \frac{\bigtriangledown U(Z,D;\alpha)}{U(Z,D;\alpha)} + \frac{\bigtriangledown \sum_D U(Z,D;\alpha)}{\sum_D U(Z,D;\alpha)}.$$

The $\frac{\bigtriangledown \sum_D U(D,Z^t,\alpha)}{\sum_D U(D1,Z^t;\alpha)}$ term presents a problem. However, from the previous iteration of the algorithm we already have a sample $D_1 \dots D_n$ from the distribution of D implied by the current value Z^t of Z. This allows us to form the following approximation:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\bigtriangledown U(D_i,Z^t;\alpha)}{U(D_i,Z^t;\alpha)} \xrightarrow{n\to\infty} \sum_D \frac{\bigtriangledown U(D,Z^t;\alpha)}{U(D,Z^t;\alpha)} P(D,Z^t;\alpha) \tag{5.3}$$

$$= \frac{\sum_D \bigtriangledown U(D,Z^t,\alpha)}{\sum_D U(D,Z^t;\alpha)} \tag{5.4}$$

$$= \frac{\bigtriangledown \sum_D U(D,Z^t,\alpha)}{\sum_D U(D,Z^t;\alpha)}, \tag{5.5}$$

so overall we have

$$\frac{\bigtriangledown U(D,Z^t;\alpha)}{U(D,Z^t;\alpha)} + \frac{1}{n}\sum_{i=1}^{n} \frac{\bigtriangledown U(D_i,Z^t;\alpha)}{U(D_i,Z^t;\alpha)} \xrightarrow{n\to\infty} \bigtriangledown \log(P(D|Z;\alpha)),$$

which may be used to form an approximation for the Langevin proposal for Z. In fact, a MALA algorithm which includes an approximation formulated in such a way is equivalent to the 'Noisy MALA-exchange' algorithm described by Alquier et al. (2016), who give a proof that the chains generated converge to the correct target distribution.

5.2 Simulation study

We conduct an experiment to compare the performance of models that implement the combination utility, with those which implement the multinomial utility. We generate 50 realisations of Gaussian random fields Z over 15×15 grids over the unit square, with $n = 30$ sampling points chosen via a utility function, at which measurements Y were taken, with Gaussian random noise on top of the true value X. The following parameter values and distributions were used.

$$Z, X|\theta, \sigma^2, \varphi \sim \text{MVN}_{N+n}(\theta, C),$$

where $C_{ij} = \sigma^2 \exp(-\frac{M^*_{ij}}{\varphi})$, and M^* is the Euclidean distance matrix pertaining to all points of interest, i.e. the focus points of the cells (at which we predict Z) and the sampling sites (at which we predict X).

$$Y|X, \tau^2 \sim \text{MVN}_n(X, \tau^2 I)$$

$$D|Z, \alpha, \beta \sim \frac{U(D,Z;\alpha,\beta)}{K(Z;\alpha,\beta)}.$$

Parameter	Actual value	Combination model mean (s.d.)	Multinomial Model mean (s.d.)
$\hat{\alpha}$	4.50	3.82 (0.954)	0.625 (0.189)
$\hat{\beta}$	1375	1570 (681)	NA
$\hat{\sigma^2}$	4.00	8.00 (0.995)	7.50 (0.934)
$\widehat{\log(\varphi)}$	1.00	2.35 (0.0456)	2.36 (0.0470)
$\hat{\tau^2}$	0.01	0.00913 (0.00188)	0.0118 (0.0132)
$\hat{\theta}$	5.00	5.09 (2.43)	5.48 (2.14)
$SSE(Z)$		32.3 (8.13)	38.8 (10.5)

Table 5.1: Mean parameter estimates, over 50 simulations, and standard deviations of estimates, and sum of squared errors when data is generated from combination utility.

As usual, the sampling locations within $\tilde{D}$ are then chosen uniformly from within their allocated cells. The parameter values are set to $\alpha = 4.5$, $\beta = 1375$, $\tau^2 = 0.01$, $\sigma^2 = 4$, $\log(\varphi) = 1$, $\theta = 5$. Using the measurements Y at the design locations $\tilde{D}$, we attempt to reconstruct Z using the hierarchical model with the above and following dependencies:

$$\varphi \sim IG(10, 500)$$

$$\sigma^2 \sim IG(60, 100)$$

$$\tau^2 \sim IG(2, 0.01)$$

$$\theta \sim N(\bar{Y}, 20)$$

$$\alpha \sim N(0, 1)$$

$$\beta \sim N(0, 2000).$$

We consider two cases for comparison, the first when $U(D, Z; \alpha, \beta)$ is the combination utility (3.8) and the second is the multinomial utility (2.1). The models are fitted using 2500 MCMC iterations, with MALA sampling for Z. The ratio of normalising constants is estimated at each iteration with 1500 design samples generated by one-step Metropolis-Hastings point swapping. In order to compare the results of the two models, we look at the mean sum of squared errors of Z across the 50 different Gaussian random fields, along with the mean parameter estimates. The results are displayed in Table 5.1. The full model fit gave a lower sum of squared errors for Z 45 times out of 50. These better results for the full combination model are due to the fact that when the preference for coverage is also taken into account, the level of preference for high values may be more accurately predicted. The preference parameters α and β were predicted reasonably well in the combination models with the average value of α slightly underestimated and β slightly overestimated. As we would expect, α was, on average, highly underestimated in the multinomial models: the extent of the preferential sampling was underestimated as the model could not account for the fact that a high preference was needed to overcome the preference for space-filling. This, in turn, has led to greater error in Z. Predictions for other parameters were largely the same in both models, σ^2 and φ were overestimated, possibly due, again, to non-identifiability

Parameter	Actual value	Combination model mean (sd)	Multinomial Model mean (sd)
$\hat{\alpha}$	1.50	2.43 (1.27)	1.86 (0.466)
$\hat{\beta}$	0.00	-0.649 (0.934)	NA
$\hat{\sigma^2}$	4.00	7.67 (1.34)	6.88 (1.45)
$\widehat{\log(\varphi)}$	1.00	2.52 (0.0731)	2.52 (0.0515)
$\hat{\tau^2}$	0.01	0.00954 (0.00211)	0.00935 (0.00207)
$\hat{\theta}$	5.00	5.90 (1.84)	5.97 (2.11)
$SSE(Z)$		128 (92.3)	132 (99.07)

Table 5.2: Mean parameter estimates for the simulation study of 50 Gaussian random fields, and sum of squared errors when data is generated from the multinomial utility.

issues as in Christensen et al. (2006).

Naturally, it is somewhat unsurprising that the true model should give a better fit. We repeat the experiment, but this time with data generated from the multinomial utility with $\alpha = 1.5$. The results are shown in Table 5.2. In this case the combination utility model fit gave a lower sum of squared errors for Z 24 times out of 50, as opposed to 26 for the multinomial model. Likewise, the space-filling term of the combination model has been predicted to be close to zero, as we would hope. The multinomial model (as we would expect) has better predictions for the strength of preference α, yet this does not appear to have had any significant detrimental effect on the prediction of Z. Other parameter value estimates are largely the same. These results demonstrate the usefulness of the combination model in giving more accurate predictions for both the strength of preference for higher values, and consequently the underlying field when there is a preference for good coverage of the region, but also that such a model does not lead to any significant inaccuracies in prediction when there is no such preference.

5.3 Scottish field ammonia data

We consider a data set comprised of 61 measurements of ammonia concentrations, taken in Autumn 2012 from a field, known as 'Corner field', with sheep in it, in Midlothian, Scotland, taken from a study of soil measurements from a mixed livestock farm in central Scotland (2012-2013), which can be found in Cowan (2019). An aerial photograph of the field in question, and the field boundary with sampling locations shown in Figure 5.2. We transform the data using the transformation $Y = 5 + \log(\text{concentrations})$ (this +5 transformation was chosen so as to make all the Y values positive, so as to make the relative strengths of preference for different sites easier to conceptualise).

Clearly, from Figure 5.2 we can see that there has been an attempt to sample relatively evenly over the area of the field, shown by the grid-like structure of sampling points. Nonetheless, the clusters in the top-right corner of the field, and at the top of the field close to the centre give some indication of a disproportionately high sampling density in areas of the field where high measurements have been taken. This suggests that a model with a utility that allows for preference for high values, and good coverage for the region might be appropriate. We seek to predict the ammonia concentrations at a set of regular grid point over whole field. We use the same transformation as above. We select these grid

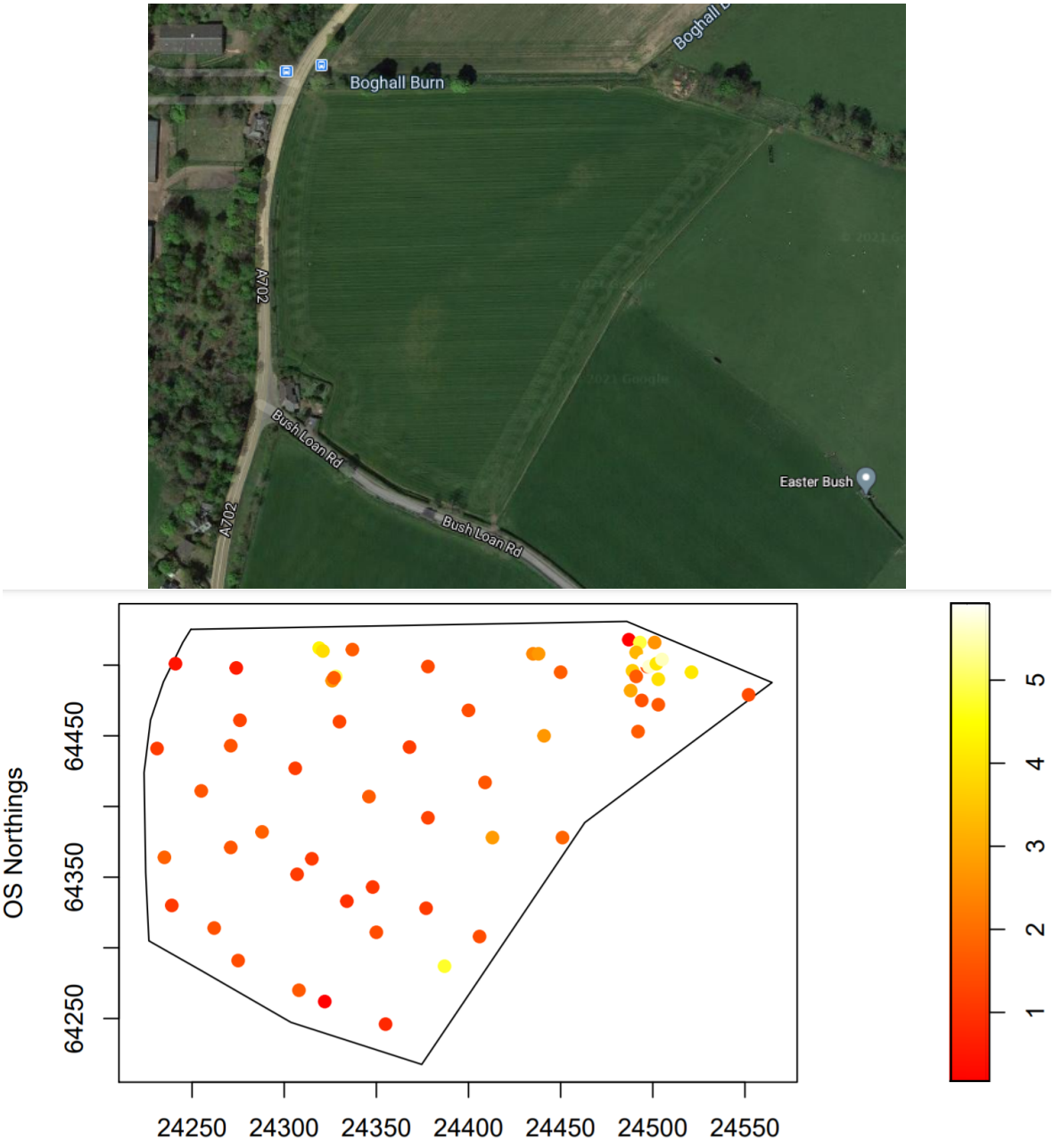


Figure 5.2: Upper: Aerial photograph of Corner Field
Imagery ©2021 CNES/Airbus, Getmapping PLC, Maxar Technologies, the Geoinformation Group, map data ©2020 United Kingdom. Lower: Corner field boundary, with sampled locations, coloured according to 5 plus the log measurements of ammonia recorded at them.

	Combination Utility mean (sd)	Multinomial Utility mean (sd)	Uniform Utility mean (sd)
$\hat{\alpha}$	1.08 (0.92)	0.869 (0.875)	NA
$\hat{\beta}$	23200 (18180)	NA	NA
$\hat{\sigma^2}$	0.893 (0.127)	0.934 (0.146)	1.33 (0.260)
$\hat{\log(\varphi)}$	-6.45 (0.203)	-6.78 (0.239)	-6.46 (0.232)
$\hat{\tau^2}$	0.0397(0.0352)	0.0386 (0.0325)	0.0400 (0.0353)
$\hat{\theta}$	1.37 (0.212)	1.43 (0.198)	1.91 (0.288)
$\bar{Z}_{\text{sampled}}$	2.33	2.31	2.20
$\bar{Z}_{\text{unsampled}}$	1.33	1.39	1.73

Table 5.3: Sheep field ammonia example: parameter estimates for the three models. $\bar{Z}_{\text{sampled}}$ denotes the mean Z values pertaining to cells from which samples had been taken, $\bar{Z}_{\text{unsampled}}$ denotes the mean Z values pertaining to cells from which samples had not been taken.

points by taking a 50×50 square grid from the extremes of the field and deleting the points which do not lie within the boundary, leaving 1526 points. Taking $\tilde{D}$ as the design, D the regularised design of monitor counts per cell, measurement values Y with regular Gaussian random field values Z, sampled point Gaussian random field values X, we assume the following model:

$$Y|X, \tau \sim N_n(X, \tau^2 I)$$

$$Z, X|\sigma^2, \varphi \sim N_{N+n}(\theta, C)$$

With $C = \sigma^2 \exp(-\frac{M^*}{\varphi})$ where M^* is the Euclidean distance matrix of all points at which Z and X are predicted.

$$\theta \sim N(2.19, 5)$$

$$\sigma^2 \sim IG(5, 0.5)$$

$$\varphi \sim IG(2.25, 0.005)$$

$$\tau^2 \sim IG(2.25, 0.0625)$$

$$\alpha \sim N(0, 100)$$

$$\beta \sim N(0, 2000000)$$

$$P(D|Z, \alpha, \beta) = \frac{U(D, Z, \alpha, \beta)}{K(Z, \alpha, \beta)}.$$

We fit three models, the first in which the utility function is the combination utility (3.8), the second in which the utility is the multinomial utility (2.1), and the third in which the utility is simply the uniform distribution, i.e. it is assumed that there has been no preferential sampling. The models are fitted via MCMC with 10000 iterations, discarding the first 2500, with the normalising constant ratio estimated, where necessary, with 1000 design samples at each iteration, the first 500 of which are discarded. The results are displayed in Table 5.3.

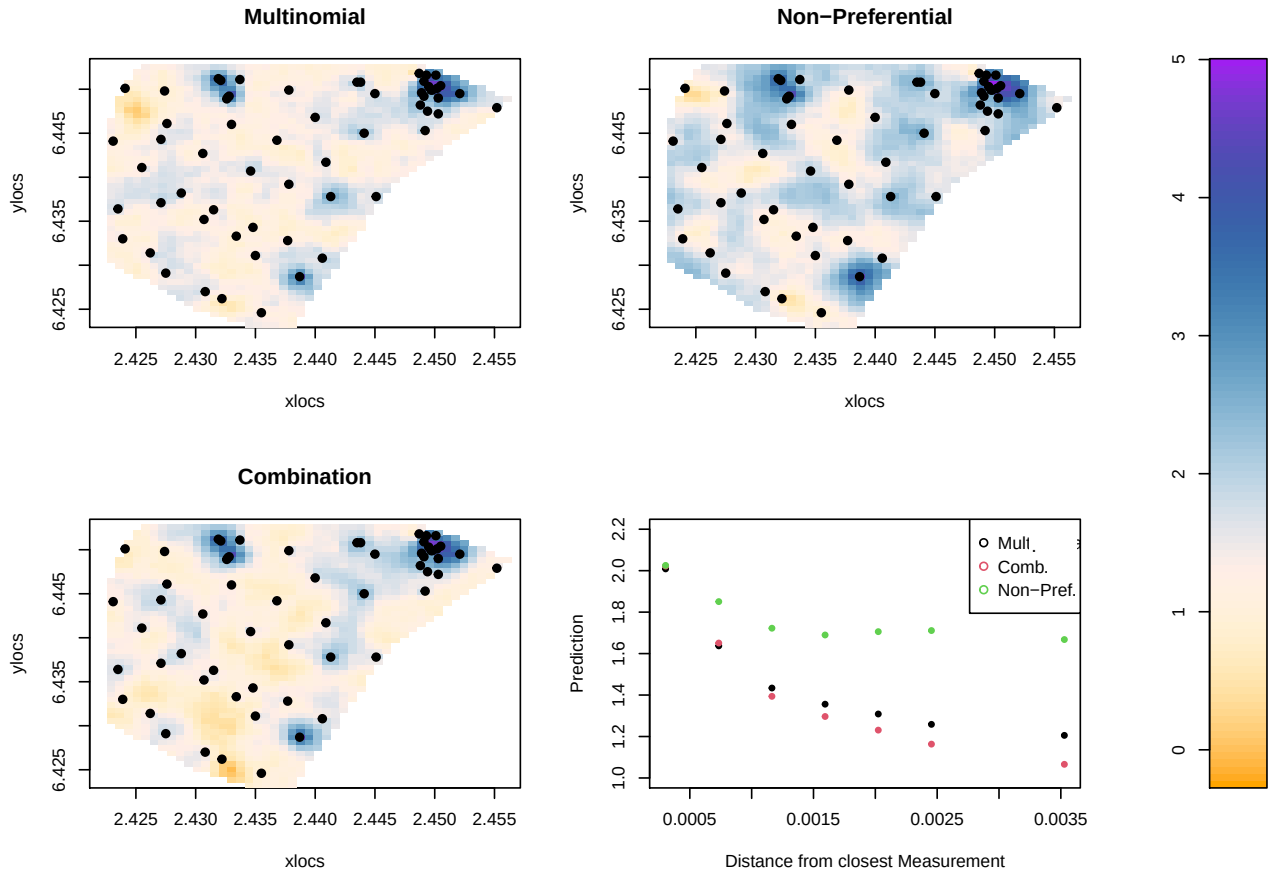


Figure 5.3: Plots showing the resulting Gaussian random fields predicted by the three different models, along with a plot of the mean predicted values of Z against distance from the nearest monitoring station.

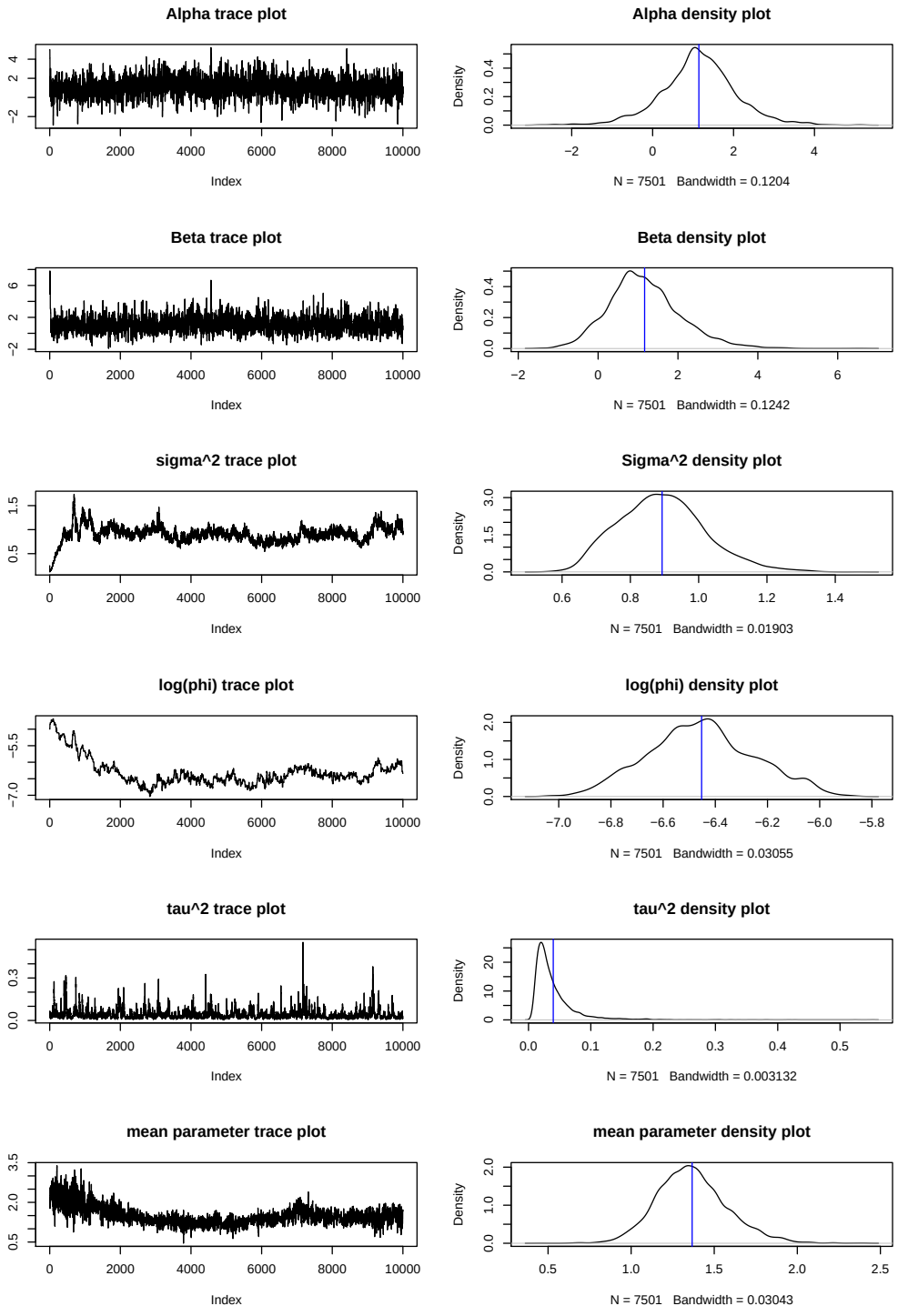


Figure 5.4: Parameter plots for the Scottish field example: Combination utility case.

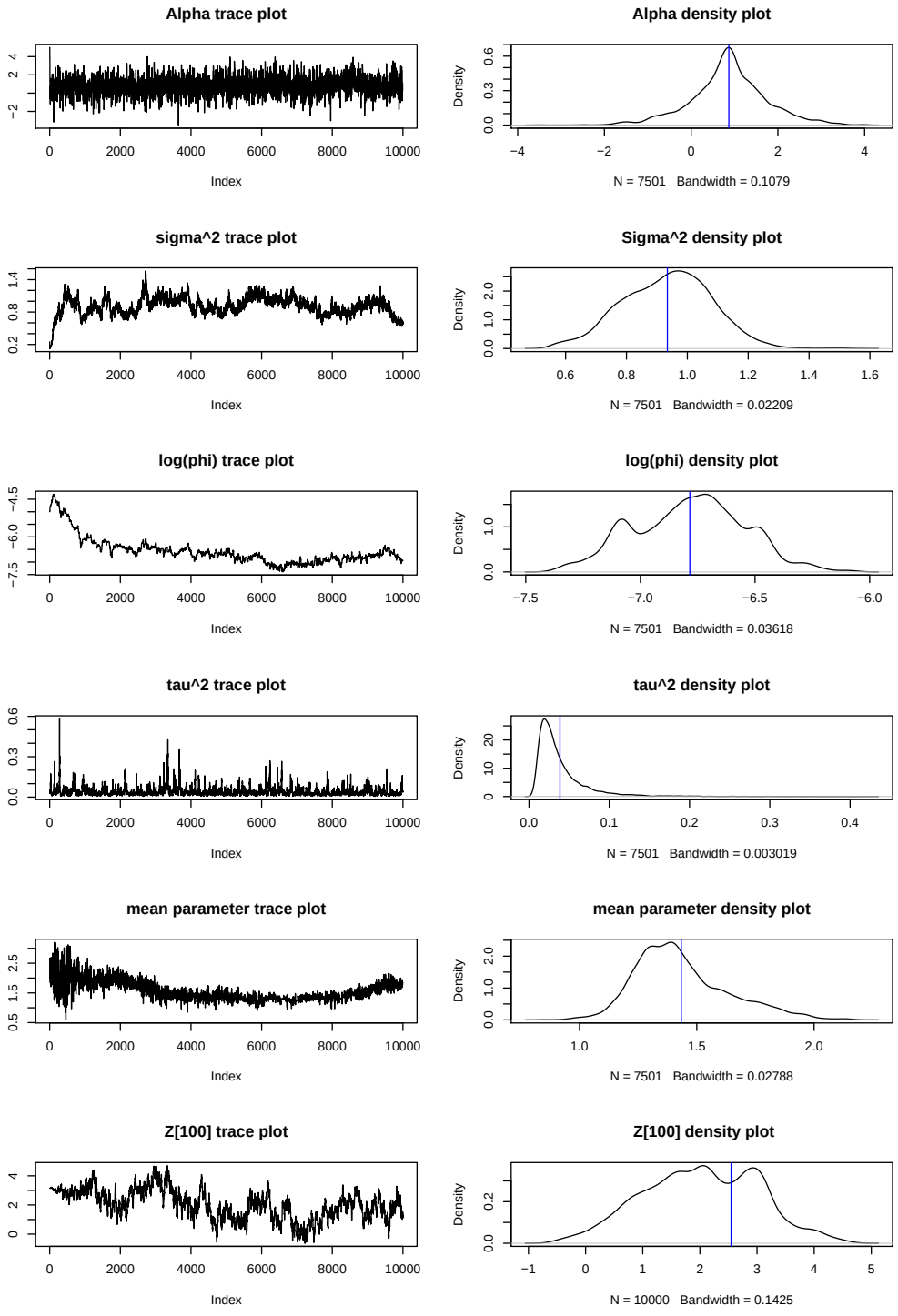


Figure 5.5: Parameter plots for the Scottish field example: multinomial utility case.

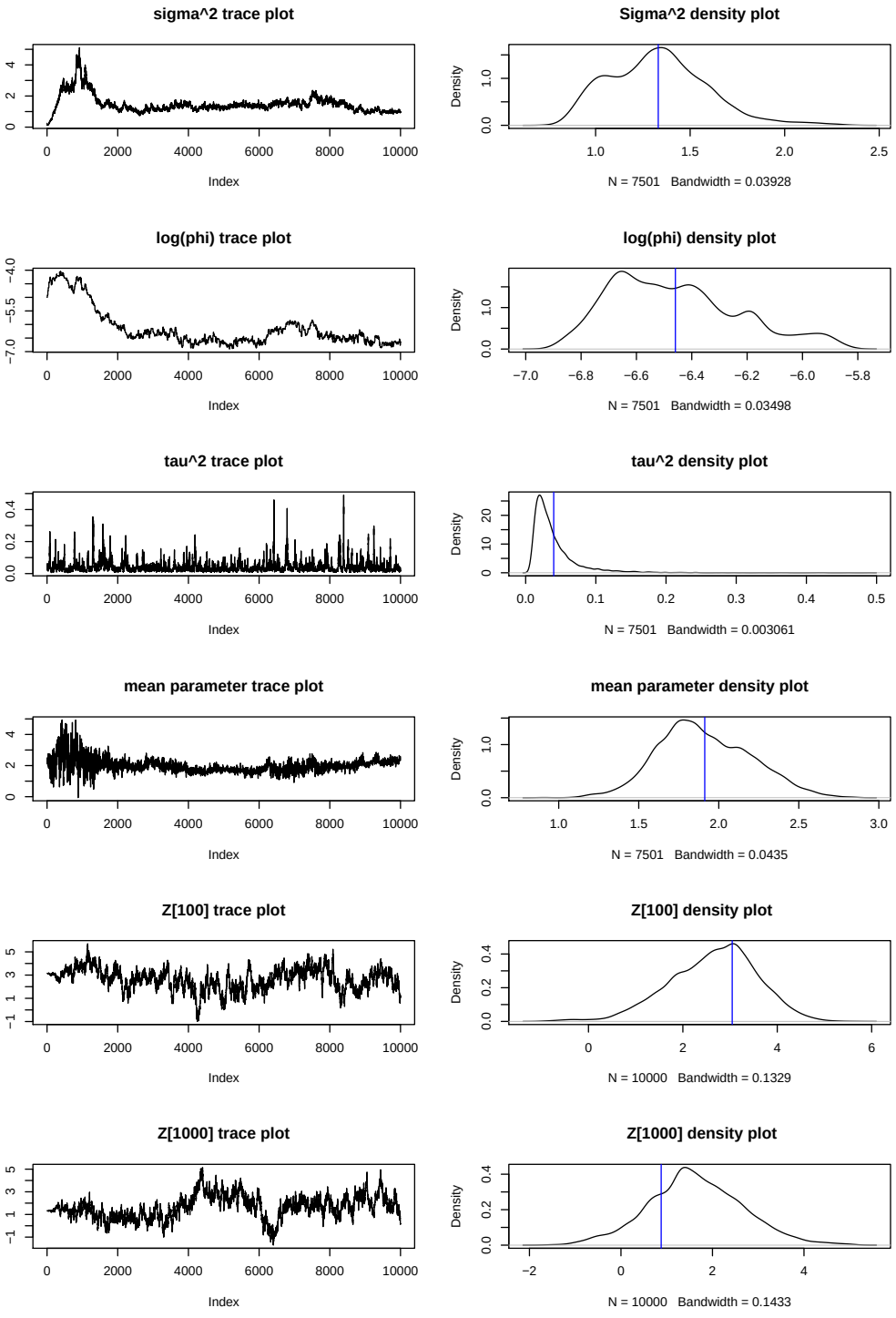


Figure 5.6: Parameter plots for the Scottish field example: non-preferential case.

The positive estimated value of the space-filling coefficient β confirms the suggestion of the grid-like structure of the monitor locations, that there has been a preference to fill the space evenly, balanced with a preference for sampling at the higher-valued locations. The knock-on effect of this is that α has been predicted more highly in the combination model than the multinomial model, and consequently the less sampled locations have been predicted to take lower values of Z, as can be seen in the first three plots of Figure 5.3. The bottom right-hand plot demonstrates this further: in general, where preferential sampling is taken into account, the further a point is from a monitor, the lower the predicted value. This effect is stronger for the combination utility, for which the strength of preference for high values is not masked by a preference for space-filling. In all three cases the parameter estimates for τ^2, $\log(\varphi)$ and σ^2 are similar, while the estimate for θ is smallest for the combination utility case, followed by the multinomial case. It is worth noting that the very low values for $\widehat{\log(\varphi)}$ is in part due to the fact that, for computational reasons, the distances (in terms of differences of Eastings and Northings) were scaled by a factor of 10000: a value of $\log(\varphi) = -6.45$ corresponds (roughly) to a correlation of 0.7 between neighbouring cells. The differences in $\hat{\theta}$ are as we would expect, as the lesser-sampled regions have been predicted to be lower. Parameter plots for the three models are shown in Figures 5.4, to 5.6.

Chapter 6

The importance of the discretisation

6.1 Discretising the space

In the preceding sections, we assume a fixed and known discretisation of the region in question, determining the domain of the utility function, and thus the probability distribution of the design. For example, we divide the region into an arbitrary number of cells of equal area, and assume that the number of monitors placed within each cell is stochastically dependent on the underlying process of interest Z only via its value at a focus point within it. In this section we study the impact of assuming such a fixed grid, and propose several strategies for selecting an appropriate discretisation. We will firstly consider the impact upon the design distribution that merging neighbouring cells has, in the particular case when a multinomial utility function is used. Next, we will propose and demonstrate how the Deviance Information Criteria (Spiegelhalter et al., 2002) may be employed to make a selection between possible candidate regular discretisations of varying coarseness. Finally, we shall explore a method, using discretisations based on Voronoi diagrams, which allows for irregular discretisations.

6.2 The Kullback Leibler divergence

We begin with an illustrative example, in which we consider the symmetric Kullback Leibler divergence (Kullback and Leibler (1951)) between two design densities implied by multinomial utility functions (2.1) and different discretisations of the region: a coarse and a fine discretisation. Given Kullback Leibler divergences $D_{KL}(P||Q)$ from distribution $P(x)$ to distribution $Q(x)$ (which are defined on the same probability space), defined to be

$$D_{KL}(P||Q) = E_{x\sim P(x)}\left(\log\left(\frac{P(x)}{Q(x)}\right)\right),$$

the symmetric Kullback Leibler divergence is given by

$$SKLD(P,Q) = D_{KL}(P||Q) + D_{KL}(Q||P).$$

We consider the probability density function for designs $\tilde{D}$ on the continuous level, given the con-

tinuous spatial process $\tilde{Z}$ i.e. we are interested in

$$p(\tilde{D}|\tilde{Z},\alpha) = \sum_D p(\tilde{D}|D)P(D|\tilde{Z},\alpha) = \sum_D p(\tilde{D}|D)\frac{U(D,Z;\alpha)}{K(Z;\alpha)}.$$

Say we have discretisation G which has N_G cells c_i, each with area A_i^G $i = 1,\ldots,N_G$. The number of monitors in cell i, in this discretisation is d_i^G. Given a discrete level design D^G (i.e. the cell locations of the monitors, but not the locations within those cells), where the superscript G denotes the discretisation, we assume that the exact locations are sampled uniformly, leading to

$$p_G(\tilde{D}|D^G) = \begin{cases} \frac{\prod_{i=1}^{N_G} d_i^G!}{\prod_{i=1}^{N_G}(A_i^G)^{d_i^G}} & \text{if } D \in D^G \\ 0 & \text{otherwise,} \end{cases}$$

which follows from the fact that, given d_i monitors in cell c_i the probability that any particular set of d_i locations is chosen will be $(\frac{1}{A_i})^{d_i}$, with $d_i!$ ways of arranging the monitors into them. Meanwhile

$$P_G(D|\tilde{Z},\alpha) = \frac{U_G(D^G,Z^G;\alpha)}{K_G(Z^G;\alpha)},$$

as usual (but with G subscripts and superscripts denoting the discretisation being used). Thus, for discretisation G we have

$$\begin{aligned} p_G(\tilde{D}|\tilde{Z},\alpha) &= \sum_{D^G} p_G(\tilde{D}|D^G)P_G(D^G|\tilde{Z},\alpha) \\ &= \sum_{D^G} p_G(\tilde{D}|D^G)\frac{U_G(D^G,Z^G;\alpha)}{K_G(Z^G;\alpha)} \\ &= m_G\left(D^G\right)\frac{U_G(D^G,Z^G;\alpha)}{K_G(Z^G;\alpha)}, \end{aligned}$$

where

$$m_G(D^G) = \frac{\prod_{i=1}^{N_G} d_i^G!}{\prod_{i=1}^{N_G}(A_i^G)^{d_i^G}}.$$

We consider the case in which the two discretisations are the same, apart from the fact that each coarse discretisation cell is made up of two adjacent fine discretisation cells. We shall here take the 'focus point' values Z^G of $\tilde{Z}$ used to be the mean over their corresponding cells, so that the values associated with the coarse discretisation cells are the means of the values of $\tilde{Z}$ associated with the associated cells in the fine discretisation. We consider a coarse discretisation C and distribution $p_C(\tilde{D}|\tilde{Z},\alpha)$ for the design, and a fine discretisation F and corresponding distribution $p_F(\tilde{D}|\tilde{Z},\alpha)$. It can be shown, using the fact that the normalising constants of these distributions do not depend on the designs D or $\tilde{D}$ that

$$SKLD(p_F(\tilde{D}|\tilde{Z},\alpha), p_C(\tilde{D}|\tilde{Z},\alpha)) = E_{\tilde{D}\sim p_F}\Bigg(\log(m_F(D^F)) + \log(U_F(D^F, Z^F;\alpha)) \tag{6.1}$$

$$- \log(m_C(D^C)) - \log(U_C(D^C, Z^C;\alpha))\Bigg) \tag{6.2}$$

$$+ E_{\tilde{D}\sim p_C}\Bigg(\log(m_C(D^C)) + \log(U_C(D^C, Z^C;\alpha)) \tag{6.3}$$

$$- \log(m_F(D^F)) - \log(U_F(D^F, Z^C;\alpha))\Bigg), \tag{6.4}$$

In the case of the multinomial utility we have, up to constants in $\tilde{D}$

$$\log\left(U_G(D^G, Z^G;\alpha)\right) = -\log\left(\prod_{i=1}^{N_G} d_i^G!\right) + \alpha \sum_{i=1}^{N_G} z_i^G d_i^G.$$

This assumes a form of the multinomial utility in which the multiplicative terms which do not depend on the design are considered to be the normalising constant. For the terms pertaining to the conditional probabilities for the continuous designs we have

$$\log\left(m_G(D^G)\right) = \log\left(\prod_{i=1}^{N_G} d_i^G!\right) - \log\left(\prod_{i=1}^{N_G} (A_i^G)^{d_i^G}\right).$$

For this multinomial utility, if we now include the assumption that, within the two discretisations, the cells all have the same area, then the area-related terms become independent of the design. This means that the m_C and m_F terms, and the factorial terms of the utility functions all cancel one another out. This leaves us with

$$SKLD = E_{\tilde{D}\sim p_F}\left(\alpha(\sum_{i=1}^{N_F} z_i^F d_i^F - \sum_{i=1}^{N_C} z_i^C d_i^C)\right) \tag{6.5}$$

$$+ E_{\tilde{D}\sim p_C}\left(\alpha(\sum_{i=1}^{N_C} z_i^C d_i^C - \sum_{i=1}^{N_F} z_i^F d_i^F)\right). \tag{6.6}$$

We first consider the second term of this. We use the assumption that the fine discretisation is the same as the coarse discretisation, with every cell divided into two equally sized parts, with the value of the element of Z associated with a cell in the coarse design the mean of the two elements of Z associated with the two merged cells from in the fine discretisation. This allows us to redefine the z^F values by associating every pair with their corresponding value: $z_i^C = z_i,\ i = 1 \ldots N_C$, in the coarse discretisation (C superscript dropped to simplify the notation), denoting them as $\{(z_1^1, z_1^2), \ldots, (z_{N_C}^1, z_{N_C}^2)\}$, with

$$z_i^1 = z_i + \epsilon_i,$$

$$z_i^2 = z_i - \epsilon_i,$$

where, without loss of generality, $\epsilon_i \geqslant 0$, meanwhile, $d_i^1 + d_i^2 = d_i$ denotes, on the left hand side, the

counts in the cells of the fine discretisation, and on the right hand side, the counts in the merged cell of the coarse discretisation. Using this notation, and linearity of expectation allows us to write

$$E_{\tilde{D}\sim p_C}\left(\alpha(\sum_{i=1}^{N_C} z_i^C d_i^C - \sum_{i=1}^{N_F} z_i^F d_i^F) - \right) = \alpha\Bigg(\sum_{i=1}^{N_C} z_i E_{\tilde{D}\sim p_C}(d_i) \\ - \sum_{i=1}^{N_C} z_i^1 E_{\tilde{D}\sim p_C}(d_i^1) + z_i^2 E_{\tilde{D}\sim p_C}(d_i^2)\Bigg).$$

Using that once the number of monitors in each cell has been chosen, the sampling locations within the cells are sampled uniformly, giving $E_{\tilde{D}\sim p_C}(d_i^1) = E_{\tilde{D}\sim p_C}(d_i^2)$ we now have that

$$\begin{aligned}
\text{LHS} &= \alpha\Bigg(\sum_{i=1}^{N_C} z_i E_{\tilde{D}\sim p_C}(d_i) - \sum_{i=1}^{N_C} (z_i+\epsilon_i) E_{\tilde{D}\sim p_C}(d_i^1) + (z_i-\epsilon_i) E_{\tilde{D}\sim p_C}(d_i^2)\Bigg) \\
&= \alpha\Bigg(\sum_{i=1}^{N_C} z_i E_{\tilde{D}\sim p_C}(d_i) - \sum_{i=1}^{N_C} z_i E_{\tilde{D}\sim p_C}(d_i^1 + d_i^2) - \sum_{i=1}^{N_C} \epsilon_i E_{\tilde{D}\sim p_C}(d_i^1 - d_i^2)\Bigg) \\
&= -\alpha \sum_{i=1}^{N_C} \epsilon_i E_{\tilde{D}\sim p_C}(d_i^1 - d_i^2) \\
&= 0.
\end{aligned}$$

Consequently, we only need the first term of (6.5), and thus, adopting the same notation, have

$$\begin{aligned}
SKLD &= E_{\tilde{D}\sim p_F}\left(\alpha(\sum_{i=1}^{N_F} z_i^F d_i^F - \sum_{i=1}^{N_C} z_i^C d_i^C)\right) \\
&= \alpha E_{\tilde{D}\sim p_F}\left(\sum_{i=1}^{N_C}((z_i+\epsilon_i)d_i^1 + (z_i-\epsilon_i)d_i^2) - \sum_{i=1}^{N_C} z_i d_i)\right) \\
&= \alpha\left(\sum_{i=1}^{N_C} \epsilon_i (E_{\tilde{D}\sim p_F}(d_i^1) - E_{\tilde{D}\sim p_F}(d_i^2))\right).
\end{aligned}$$

By properties of the multinomial distribution we have

$$E_{\tilde{D}\sim p_F}(d_i^1) = \frac{n\exp(\alpha z_i^1)}{\sum_{j=1}^{N_c}\exp(\alpha z_j^1) + \exp(\alpha z_j^2)},$$

Substituting this in gives us

$$\begin{aligned} SKLD &= \frac{n\alpha \sum_{i=1}^{N_c} \epsilon_i (\exp(\alpha z_i^1) - \exp(\alpha z_i^2))}{\sum_{j=1}^{N_c} \exp(\alpha z_j^1) + \exp(\alpha z_j^2)} \\ &= \frac{n\alpha \sum_{i=1}^{N_c} \epsilon_i (\exp(\alpha(z_i + \epsilon_i)) - \exp(\alpha(z_i - \epsilon_i))}{\sum_{j=1}^{N_c} \exp(\alpha(z_j + \epsilon_j)) + \exp(\alpha(z_j - \epsilon_j))} \\ &= \frac{n\alpha \sum_{i=1}^{N_c} \epsilon_i (\exp(\alpha z_i)(\exp(\alpha\epsilon_i) - \exp(\alpha\epsilon_i)))}{\sum_{j=1}^{N_c} \exp(\alpha z_j)(\exp(\alpha\epsilon_j) + \exp(-\alpha\epsilon_j))}. \end{aligned}$$

We can see several things from the above. Firstly, the distance between design distributions is minimised when each ϵ_i is as small as possible, suggesting that a merge or split of cells makes little difference in relatively homogeneous areas. Secondly, this difference is less significant for the lower-valued cells (assuming $\alpha > 0$). Whilst, without knowledge of Z, this does not imply a definitive method for selecting a discretisation, it suggests the principle that merging areas corresponding to very different Z values into one 'zone' may have an impact on the design distribution, consequently making preferential sampling less detectable. Likewise, it suggests that finer grids may be more appropriate where the scale of the variation is smaller.

6.2.1 Grid selection using deviance information criteria

Having seen that the choice of discretisation can make a difference to the model where cells contain highly varied values of the process of interest, we seek a method for selecting an appropriate fineness of discretisation. We consider an approach of fitting a selection of models with various different grid sizes, before comparing them using model selection criteria, such as the Deviance Information Criteria (Spiegelhalter et al. (2002)), which are a generalisation of the Akaike Information Criteria, useful for selection of Bayesian hierarchical models. The deviance is defined as

$$\mathcal{D}(\theta) = -2\log(p(y|\theta)) + C,$$

where y is the observed data, θ the underlying parameters, and C a constant which cancels out in comparisons, and may thus be ignored. Using this definition of deviance, the deviance information criterion is then defined to be

$$\text{DIC} = \overline{\mathcal{D}(\theta)} + P_d,$$

with

$$P_d = \overline{\mathcal{D}(\theta)} - \mathcal{D}(\bar{\theta}).$$

This lends itself easily to MCMC model fitting as $\mathcal{D}(\theta)$ can be calculated using samples of the chain. In our case we have

$$\mathcal{D}(\tilde{D}, Y|Z, X, \tau^2, \varphi, \sigma^2, \theta) = -2\log(P(Y|X)P(\tilde{D}|Z, \alpha)).$$

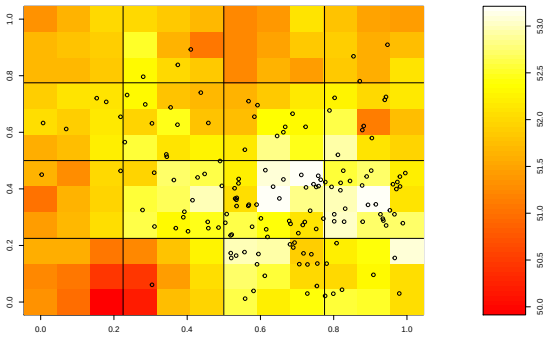


Figure 6.1: 12×12 realisation of a Gaussian random field, with superimposed sampling locations based on a multinomial utility and a 4×4 discretisation of the region.

Thus we have

$$\mathcal{D}(\bar{\theta}) = -2\log(P(Y|\hat{X})P(\tilde{D}|\hat{Z},\hat{\alpha})),$$

and

$$\overline{\mathcal{D}(\theta)} = -\frac{2}{T}\sum_{i=1}^{T}\log(P(Y|X^i)P(\tilde{D}|Z^i,\alpha^i)),$$

where $\hat{Z}$, $\hat{X}$ and $\hat{\alpha}$ are the average values of Z^i, X^i and α^i respectively, which are the values of their respective variables at step i, of the Markov Chain, of T steps. The design distribution normalising constants may be estimated as usual at each step and stored: we start our chains with $\alpha = 0$, leading to a normalising constant equal to one, which enables us to estimate single normalising constants at each step (rather than ratios) by using the estimate from the previous iteration. We carry out some simulations in order to test the appropriateness of such criteria for choosing a discretisation.

Numerical experiments

We begin with an experiment in which Z is a 12×12 realisation of a Gaussian random field over the unit square with parameters $\theta = 50$, $\sigma^2 = 5$, $\log(\varphi) = 1.5$, with an exponential covariance function. The cells in which sampling locations are to be placed are chosen using a 4×4 discretisation of the unit square and the multinomial utility i.e. the probability of any cell being selected for any particular monitor is proportional to the exponential of α (in this case $\alpha = 3$) multiplied by the mean of the nine Z values which correspond to that cell. Having chosen the cells the locations within them are sampled uniformly. This set-up is demonstrated in Figure 6.1.

We then re-fit several models to the resulting data, with the differing assumptions that the discretisation of the unit square used to select designs have been 2×2, 3×3, 4×4 , 6×6 and 12×12. The results are shown in Table 6.1. In this case this approach has selected the original, correct discretisation. While these results are favourable, we conduct further experiments to determine whether they are consistently so. As before, we generate a 12×12 realisation of a Gaussian random field on the unit square, yet each time we select the sampling points based on a 2×2, 3×3, 4×4,6×6 or 12×12 grid, selected at random. The five models with each of these grid sizes are then fitted to the generated data, and the one which

Discretisation	DIC
2×2	-1377.8
3×3	-1355.202
4×4	-1418.259
6×6	-1379.762
12×12	-1388.396

Table 6.1: DIC values for the different discretisations, with the lowest value corresponding to the original discretisation.

gives the smallest value of the deviance information criterion is selected and recorded. Having repeated this experiment 100 times, we found that the correct (original) discretisation was selected 81 times.

We repeat the same process with a smaller $\alpha = 1$. We have showed that the Kullback Leibler divergence between densities based on different discretisations is dependent on α, and so we can expect that a smaller value of α will lead to the 'correct' model being selected a smaller number of times - this is indeed the case, the correct discretisation was selected only 54 times. This, while better than the 20 times we would expect if the method were totally useless, is somewhat less promising. However, we note that with a lower value of α, a poorly specified discretisation makes less of a difference.

6.3 Treating the discretisation as a random variable

The DIC-based method above, whilst helping to encapsulate the resolution of the experimenter's knowledge of the underlying field, has the obvious limitation of dealing only in terms of fixed cells of differing size.

Suppose now that we have the situation in which the experimenter has knowledge of some systematic differences in irregular sub-regions of the region in question, such as historical land use: previous use of a particular fertiliser on one half of a field but not on the other, a certain area used as for waste disposal in previous years etc.. Likewise, there could be differing local regulations of various authorities of which we, as the modeller, have little knowledge. Suppose also that the number of monitors assigned to one of these sub-regions is dependent on the mean value of the underlying field within them. We consider the possibility that it may be misleading to use a model in which the design distribution depends on the values close to the monitors at much higher resolution than that of any prior knowledge of the experimenter, especially in cases in which there is low spatial correlation. Assuming we have little knowledge of the sub-regions in question, we seek to build a model that allows for this possibility, and explore the idea that the discretisation itself may be treated as a random variable to be predicted from the data.

We implement this random discretisation assumption by assuming that the discretisation takes the form of a Voronoi diagram, a partition of a region defined by a set of 'centres'. Any point within the region belongs to the partition element associated with its closest centre. This can be visualised in Figure 6.2. In our case the defining 'centres' of the partition may be located at any of the N 'fine discretisation' points at which we are predicting the values of the Gaussian random field Z. The location

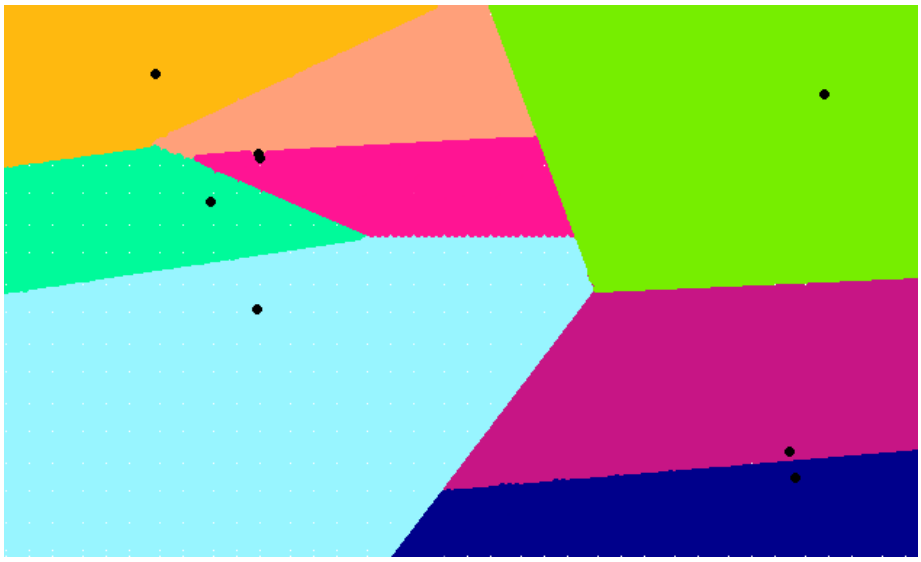


Figure 6.2: Voronoi diagram on the unit square with eight Voronoi centres. The different colours correspond to the different Voronoi cells, or 'zones'.

and number of these Voronoi centres may be sampled in a similar way to the other unknown parameters of the system. We shall focus on multinomial utilities here, as space-filling utilities make less sense where the 'cells' over which the utilities are defined are, in fact, (possibly very large) zones.

6.3.1 Putting prior distributions on cell counts

We may wish to assign a prior distribution to the number of Voronoi cells. An obvious choice would be the binomial distribution. However, if it is our intention that the number of cells should follow such a distribution (or any that we choose) we must adjust for the fact that there are $\binom{N}{k}$ distinct Voronoi diagrams with k cells, inducing a natural bias to diagrams with more cells. In order to account for this, any prior distribution on the Voronoi diagrams that will result in the number of cells following a specific distribution must include a term dividing by $\binom{N}{k}$. In practice, when the binomial distribution is used, this simply means leaving off the binomial coefficient. Care must be taken when selecting the prior mean number of cells: clearly there is a trade-off, with more cells we are more likely to find a Voronoi diagram, combinations of the cells of which form more exactly the original cells, however, with fewer cells the coefficients of the elements of Z in the Voronoi-implied model may in fact be closer to their values in the original model. Likewise, a smaller number of cells gives more easily interpretable results.

6.3.2 Sampling the Voronoi diagram

Within the wider MCMC algorithm in which we sample Z and the other parameters we sample the Voronoi diagram. We follow the method of Kim et al. (2005) who use uncertain Voronoi diagrams to fit piecewise Gaussian processes. The key difference is that we treat the sampling process as a piecewise process, rather than the underlying Gaussian random field.

The distribution of the design, given discretisation G is given by

$$P(D|Z,\alpha,G(n_G)) = \frac{U(D,Z;\alpha,G(n_G))}{K(Z,\alpha,G(n_G))}.$$

Likewise, we have a prior distribution $P(n_G)$ on the number of cells in the Voronoi diagram. The full conditional distribution on G is thus given by

$$P(G|.) = \frac{P(n_G)U(D,Z;\alpha,G(n_G))}{K(Z,\alpha,G(n_G))}.$$

The following Metropolis-Hastings algorithm enables us to sample from this distribution.
Algorithm:

1. We denote the discretisation at iteration i by G_i, with n_{G_i} Voronoi cells. Begin with n_{G_1} cells as Voronoi centres. Set the maximum possible number of Voronoi centres to be n_{max}

2. At step i, choose whether to add, remove, or change the location of a site with probabilities p_{add}, p_{remove}, and p_{move}, with

$$p_{add} + p_{remove} + p_{move} = 1$$

   If we have $n_{G_i} = n_{max}$ or $n_{G_i} = 1$ these probabilities should be altered to reflect that we can only move or delete a centre, or add a centre, respectively.

   (a) If we have chosen to **add** a Voronoi centre, randomly select (with equal probability) an as yet unselected site to be an additional Voronoi centre, to form $G^{'}$. Calculate the ratio $\frac{P(G^{'}|.)}{P(G_i|.)}$, and the transition probability ratio:

$$\frac{P(G^{'} \to G_i)}{P(G_i \to G^{'})} = \begin{cases} \frac{(N-1)p_{remove}}{2} & \text{if} \quad n_{G_i} = 1 \\ \frac{p_{remove}(N-(n_{max}-1))}{(p_{remove}+p_{move})n_{max}p_{add}} & \text{if} \quad n_{G_i} = n_{max} - 1 \\ \frac{p_{remove}(N-n_{G_i})}{(n_{G_i}+1)p_{add}} & \text{otherwise.} \end{cases}$$

   (b) If we have chosen to **remove** a site, randomly select one of the current Voronoi centres (with equal probability) for removal, to form $G^{'}$ Then calculate the ratio $\frac{P(G^{'}|.)}{P(G_i|.)}$, and the transition

probability ratio:

$$\frac{P(G^{'} \rightarrow G_i)}{P(G_i \rightarrow G^{'})} = \begin{cases} \frac{p_{add}(p_{remove}+p_{move})(n_{max})}{(N-(n_{max}-1))p_{remove}} & \text{if} \quad n_{G_i} = n_{max} \\ \frac{2}{(N-1)p_{remove}} & \text{if} \quad n_{G_i} = 2 \\ \frac{p_{add}(n_{max}-1)}{(N-(n_{max}-2))p_{remove}} & \text{if} \quad n_{G_i} = n_{max} - 1 \\ \frac{p_{add}n_{G_i}}{(N-(n_{G_i}-1))p_{remove}} & \text{otherwise.} \end{cases}$$

(c) If we have chosen to **move** a Voronoi centre we randomly select one of the current Voronoi centres to be removed, and randomly select, with equal probability, one of the unselected sites, including that which has been removed, to become a new Voronoi centre to form $G^{'}$. Then calculate the ratio $\frac{P(G^{'}|.)}{P(G_i|.)}$, and the transition probability ratio:

$$\frac{P(G^{'} \rightarrow G_i)}{P(G_i \rightarrow G^{'})} = 1.$$

3. Set $G_{i+1} = G^{'}$ with probability

$$\min\left(\frac{P(G^{'}|.)}{P(G_i|.)}\frac{P(G^{'} \rightarrow G_i)}{P(G_i \rightarrow G^{'})} \quad , 1 \quad \right),$$

else set $G_{i+1} = G_i$.

4. Repeat from step 2.

We may use this algorithm as part of the wider MCMC algorithm to sample a new Voronoi diagram after a specified number of iterations. It is worth noting that at each iteration of the above algorithm, we require the computation of the normalising constant of the utility function, meaning that the use of this method will likely be too computationally expensive to use with any utility function for which we cannot calculate the normalising constant exactly. Thus we restrict our investigations to the multinomial utility.

6.3.3 Experiments on the usefulness of this model

We conduct several experiments, with varying levels of spatial correlation, strength of preference etc. and generate preferentially-sampled data from the Voronoi-implied multinomial model. Initial explorations suggest that for a zones model to be useful we require a data set with a sufficiently large number of zones, with varying mean values of Z within them, and, in turn, a varying number of sampled points in order that α might be estimated with a reasonable level of accuracy: if we have only two or three zones containing sampling sites, it is difficult to get a good idea of the strength of preference. For this reason we define the number of zones in the simulated data to be $n_v = 18$.

We use the following model:

$$C = \sigma^2 \exp\left(\frac{-M}{\varphi}\right)$$

$$Z, X|\sigma^2, \varphi \sim N_{n+N}(0, C)$$

$$Y|X, \tau^2 \sim N_n(0, \tau^2 I)$$

$$D|Z, \alpha \sim \frac{n!}{d_1^*, \ldots d_{n_v}^*} \frac{\exp\left(\alpha \sum_{i=1}^{n_v} \left(\frac{1}{a_i^*} \sum_{j \in S_i^*} z_j\right) d_i^*\right) \prod_{i=1}^{n_v} (a_i^*)^{d_i^*}}{\left(\sum_{k=1}^{n_v} a_k^* \exp(\alpha \sum_{j \in S_k^*} z_j)\right)^n},$$

with $\sigma^2 = 1$, $\tau^2 = 0.01$. S_i^* denotes the set of fine design indices (i.e. the indices corresponding to the locations at which Z is predicted), relating to the i^{th} Voronoi zone, d_i^* the number of sampling points within it, and a_i^* its area. Values of α and φ will be varied for the different cases we shall look at. The Voronoi cells are determined by the locations of randomly selected Voronoi centres. In attempting to recover Z, we fit the following model

$$\sigma^2 \sim IG(101, 102)$$

$$\tau^2 \sim IG(0.01, 2.1)$$

$$\alpha \sim N(0, 10)$$

$$C = \sigma^2 \exp\left(\frac{-M}{\varphi}\right),$$

where M is the distance matrix relating to the locations at which Z and X are predicted.

$$Z, X|\sigma^2, \varphi \sim N_{n+N}(0, C)$$

$$Y|X, \tau^2 \sim N_n(0, \tau^2 I)$$

$$D|Z, \alpha \sim \frac{U(Z, D; \alpha)}{K(Z, \alpha)}$$

$$n_v \sim p_v^{n_v}(1 - p_v)^{(N_v - n_v)}$$

With $p_v = 18/N$ and $N_v = N$ and $n_{max} = 100$. $\log(\varphi)$ was considered fixed and known. α, $\log(\varphi)$ and n_v were varied between experiments. We used a $N = 45 \times 45$ regular discretisation of the unit square with $n = 50$ sampling points. We fit each model twice, the first with the utility function taking the functional form of that of the generating model, the second with the regular multinomial utility function on the fine grid. We calculate the sum of squared errors for both models. We carry out 20 repeats for each experiment.

Results:

1. **Experiment one:** In this experiment we have $\alpha = 2$, $\log(\varphi) = -2$, 20 repeats.

	Zones model	Regular Multinomial
SSE(Z)	1554	1583
$\hat{\alpha}$	2.16 (1.05)	0.578 (0.237)

Table 6.2: Comparison of results for zones model, compared with multinomial model.

The zones model gave a lower sum of squared errors for Z 10 out of 20 times.

2. **Experiment two:** In this experiment we have $\alpha = 2$, $\log(\varphi) = -4$, 20 repeats.

	Zones model	Regular Multinomial
SSE(Z)	1818.2	1907.118
$\hat{\alpha}$	1.38 (0.802)	0.403 (0.289)

Table 6.3: Comparison of results for zones model, compared with multinomial model.

The zones model gave a lower sum of squared errors for Z 11 out of 20 times.

These results show a somewhat disappointing lack of difference in performance between the two methods. We consider why this might be. We first consider the case in which there is low within-zone variance: a consequence of having overall high spatial correlation. This, by definition, means that the values of the field Z are close to the mean value of the zone to which they belong. This in turn brings the two design distributions, multinomial and zone-implied, close together, as the zone implied model tends to the multinomial model as the within zone variance tends to zero. In addition to bringing the design distributions closer together, a reasonably high level of spatial correlation reduces the differences in the posterior distributions for Z further: the zones model effectively 'shares out' the increase in Z implied by the number of points in the zone among the cells, whereas in the multinomial model, the increase in Z in a sampled cell also increases the values in the nearby cells, thanks to the spatial correlation in the model, resulting in a similar effect. Clearly in the cases we have experimented with, the gains from fitting the true model are too small to make any significant difference, when combined with other errors brought in by uncertainty in the zone boundaries and parameter estimation. It is worth noting that this situation may reflect the real-life situation in which the zones, such as an industrial area, or a park etc. are informally selected because of the homogeneity of activity within them. We have seen, in the first half of this chapter, with reference to the multinomial utility and the Kullback Leibler divergence, that subdividing largely homogeneous regions makes little difference.

We now consider the case in which there is large within-cell variance, and the two models are more distinct. We can split this into two cases: the case in which the sampled Y values are representative of the zone in which they are located, and the case in which they are not. We start with the case in which the Y values (or, more specifically, their mean value over the zone to which they correspond) are not very representative of their zone mean. The possibility of this happening is not unreasonable in the case when there is a high within-zone variance, especially with a small number of monitors. In this case the parameter α will be poorly estimated. Likewise, when the multinomial model is used to recover the surface, the values sampled will bear little relation to the preferences of the experimenter, rendering both models uninformative at best. In the second case, the Y values may well be representative of the

zones in which they are situated, and the value of the strength of preference parameter well estimated. However, the within-cell variability means that we do not gain a great deal of information about the specific values of Z within a zone, only their mean – the preference part of the model may just as easily reward Z values for which the high and low values of Z are switched around. In summary of these three cases, the two models, multinomial and zones, seem only to have a significant level of difference when the within cell variance is high enough to have also decreased the usefulness of both of them.

6.3.4 Future directions

While the above zones procedure, which assumes zones within a single Gaussian random field, seems to produce little advantage in terms of reducing the mean prediction error, it still may be useful to take account of preferential sampling in a situation in which the region of interest is subdivided into disjoint regions, both in terms of the underlying field Z and the sampling process, by modelling the dependence of the number of monitors per zone on the mean value of Z in that zone. This could be particularly useful

Pope et al. (2019) describe a procedure for modelling disjoint Gaussian processes over a region, thus allowing for discontinuities. This may present a more appropriate framework for modelling data in which the values of the underlying process of interest are dictated by different localised factors, leading to sharp discontinuities between zones. This also uses Voronoi tessellations to construct (not necessarily convex) sub-regions of one or more Voronoi cells. In addition to adding, moving, and deleting Voronoi centres, the change in discretisation may include the reclassification of one of the Voronoi cells, with respect to its sub-region. This is fitted in a process known as 'reversible jump MCMC'. After each proposed re-discretisation, new Gaussian process parameters are calculated, using maximum likelihood estimation, and the proposed discretisation, along with these parameters is accepted according to the Metropolis-Hastings ratio implied by the proposed and current discretisation and parameters. After this update, a Gaussian process is fitted to each of the regions. This is possible because in this scenario the Gaussian process may be integrated out, which is not the case where there is a preferential sampling component in the model. Thus a similar approach, incorporating preferential sampling may contain the following update:

1. Propose a switch.
2. For each region, sample values of σ^2, φ, θ, X, given current values of Z, τ^2 and Y.
3. Sample Z, given current values of X, σ^2 and α, the strength of preference parameter. This may involve proposing Z values from each sub-region, joining them together, and accepting or rejecting this amalgamated Z value according to the design distribution, or, alternatively, Gibbs sampling each sub-region iteratively according to the overall design distribution.
4. Repeat from step 2 for a pre-specified number of iterations, taking the mean of the parameter, X and Z samples over the iterations, in order to find values of Z which correspond to this new discretisation.

5. Accept or reject this proposed switch and its corresponding Z, X and parameters, according to the Metropolis-Hastings ratio of the posterior distribution, made up of the likelihood of the data Y and D, given these values, and any prior distribution on the discretisation.

We envisage that Voronoi diagrams and their scope for describing uncertain partitions of the space in question may also be useful in other ways. For example, there may be holes in the region (e.g. a lake), differing terrains and land usages with uncertain boundaries etc. for which there are differing costs and constraints associated with monitor placement. This could be described by a 'cost surface' made from a Voronoi diagram: those fine-discretisation cells corresponding to the same zone will have the same penalisation term within the utility function. This cost surface could be known, known locationally but with unknown parameters, partially known, or unknown. It may be possible to include prior information in the form of specified 'rewards' for adjacent cells being assigned to the same Voronoi cell.

Chapter 7

Estimating preference using multiple data sources

7.1 Why might we combine data sets?

We may find ourselves in a situation in which we have access to different sources of data relating to the process of interest, with varying levels of preference. In this case, care must be taken to assign appropriate, and possibly different, utilities to the sampling schemes relating to the different data sets. The data may take the form of multiple separately-planned sampling networks. Here, intentions of the experimenters may have been different, and to amalgamate the two networks may be inappropriate. Such action may lead to misleading inferences based on the assumption of preferences that played no part in the choice of certain sampling locations, or, conversely, underestimation of preferences, due to different preferences cancelling one another out. There are undoubtedly cases in which this should not present a problem: when the experimenters have reasonably similar preferences, for example, when considering an air pollution monitoring network in which monitoring stations are set up by many different stakeholders, some with a possible preference for higher values, it would be valid to model the monitor locations with a multinomial utility or, if the experimenters had knowledge of where the others were positioning or had positioned their stations, (assuming data sharing), a repulsive or space-filling combination utility. However, in other situations, we must take account of different preferences associated with different sampling designs for networks measuring the same quantity, by assigning different utility functions to the different sampling schemes.

Secondly, situations may arise in which, in order for accurate predictions of the level of preference to be made, other sources of data must be brought in. For example, suppose the sampling scheme were so restrictive as to only contain values from a very narrow subset of the range of the process of interest, yet, by coincidence or design, the lower values within that subset were more highly-sampled. It is easy to see here the difficulty in estimating the level of preferential sampling. In such a case the preference may be better estimated with access to other data, to which we do not apply the same utility function. We demonstrate this point with the following simple example.

Example: Preferential sampling in a highly restricted range.

A simple Gaussian random field is realised over an 18×18 regular grid on the unit square, giving focus-

point values Z. This Gaussian random field has mean $\theta = 2$, variance $\sigma^2 = 0.5$, correlation $\varphi = 0.5$ and an exponential correlation function. The first design, involving $n_1 = 30$ points is restricted to the cells associated with the highest ten percent of Z values, from which measurement sites were selected with a multinomial utility function (2.1) with $\alpha = -3$, i.e. negative preferential sampling, within this restricted range. The measurements are taken at the focus points of their cells, with mean-zero normal measurement error with variance $\tau^2 = 0.1$. The second, non-preferential, design with $n_2 = 15$ points is selected using combination utility (3.8) with $\alpha = 0$ and $\beta = 50$, to give a space-filling, grid-like design. Again, the noisy measurements were taken at the cell focus points. Z was then predicted in three ways:

1. (M1) Using only the measurements from the first design (restricted range), with an assumed multinomial utility.

2. (M2) Using the measurements from both designs, assuming a multinomial utility for the first (restricted range) design, but no preference for the second (grid-like).

3. (M3) Using measurements from both designs, but assuming no preferential sampling in either.

The models, as the utility was multinomial and did not involve an intractable normalising constant, were fitted using Hamiltonian Monte Carlo in Stan (Stan Development Team (2020)), with 5 chains of 1900 samples each, the first 400 of which were discarded as burn-in. Inverse gamma prior distributions with parameters (1,1) were imposed on both τ^2 and σ^2 and mean-zero normal prior distributions with variance of 1 were imposed on both α and θ. The predictions are shown in Figure 7.1.

For the three models the mean squared errors in Z were 0.968, 0.0793 and 0.1339 for the preferential only model (M1), the two data-set model assuming preference (M2), and for the uniform model (M3) respectively. For the two models in which preference was assumed the predicted values of $\hat{\alpha}$ were -1.94 (sd 9.29) for preferential only model (M1), and 8.02 (s.d. 2.21) for the two-data set model (M2).

There is, in this case, a clear advantage in using the method that combines data sets but assumes different sampling schemes for them, not least due to the better estimation of preference in the preferential data. Likewise, when using both data sets there is an advantage in allowing for the possibility of preferential sampling in design one, even when the wrong utility function is used (i.e. in this case a model which did not have the same functional form as the generating model).

This is clearly a worst-case scenario example, which required some degree of reverse-engineering. In practice, one would be remarkably unfortunate to come across a data set with such mis-informative preference. (Reassuringly, in many examples where measurements are only taken from a high-valued narrow subset, even with monitoring sites clustered towards the lower end of these values, the spatial correlation, combined with lower values round the edges of the highly sampled areas leads to detection of positive preference.) In practice, in a situation such as this, one would hope that even vague knowledge of the process of interest might lead to the imposition of a prior distribution favouring positive values of α. Nonetheless, this scenario highlights the rewards that can be gained from estimating the preference of one particular set of measurements using any extra data available. The point here is not simply that we get better results when we have access to more data, but that with the extra data, the preferential sampling within the original data is better recognised, and itself becomes more useful by indicating that sites far from heavily sampled areas may take lower values.

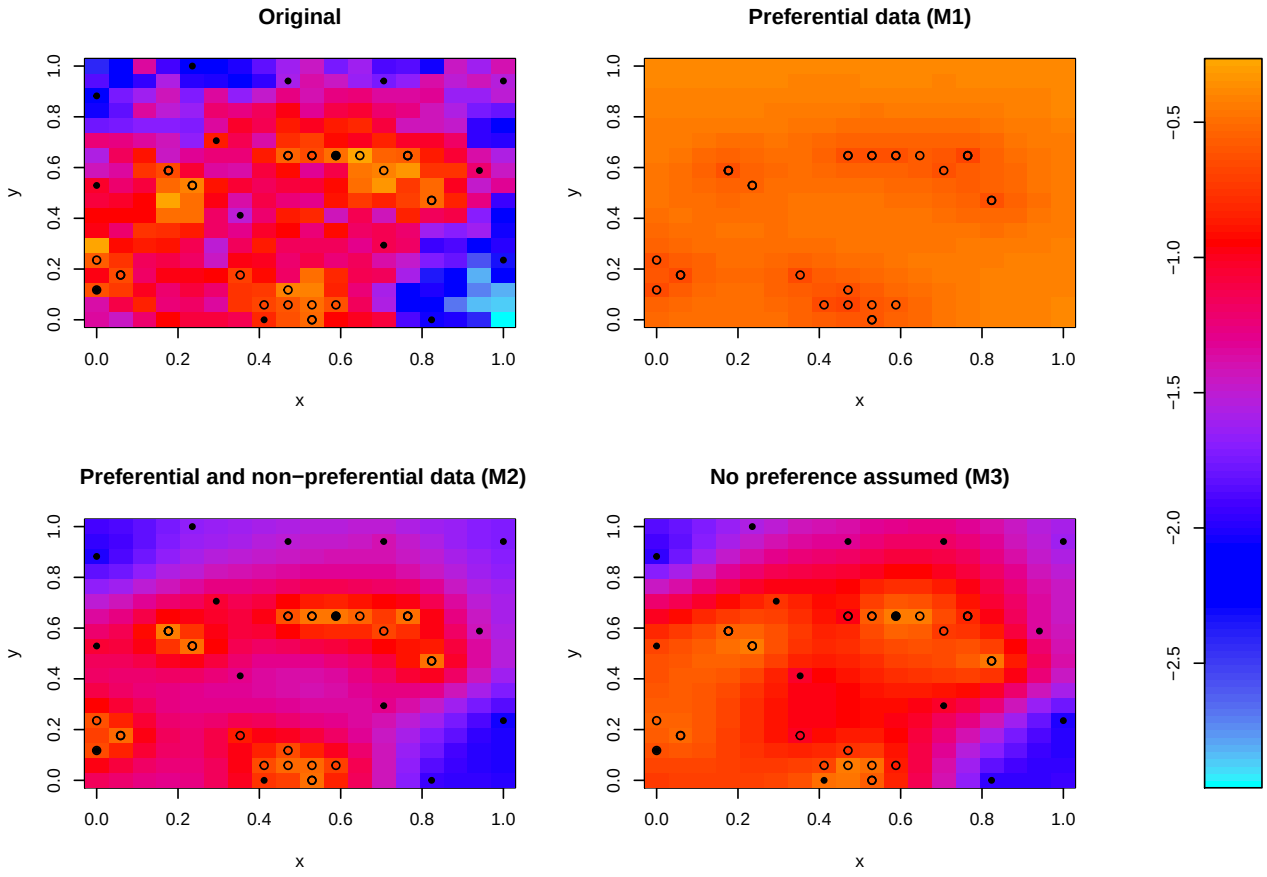


Figure 7.1: Example surface with preferential sampling locations (unfilled points) chosen with a negative preference utility function on the highest ten percent of possible sampling sites, and non-preferential sampling sites (filled points). Predicted Gaussian random fields constructed with just the preferential data, both data sets, with and without the assumption of preference in the preferential data.

Furthermore, combining data sets may also be useful when the sources are not from the same type of measurement system. For example, we may have two sources of air pollution data: one from a network of ground level monitors measuring the concentrations of a pollutant at a particular point, and the other from satellite, measuring pollution concentrations at a course resolution. For example, the Tropomi instrument, which measures terrestrial air pollution aboard the European Space Station's Copernicus Sentinel 5 satellite has a spatial resolution of 3.5×7 km (European Space Agency (2019)). Say we have ground level data

$$Y = (Y_1, \dots Y_{n_1}),$$

and satellite data

$$Y^{'} = (Y_1^{'} \dots Y_{n_2}^{'}),$$

and our aim, as usual, is to predict values $z_1, \dots z_N$ over a regular grid. Assuming this grid is aligned so as to be a finer sub-grid of that at which the satellite measurements are taken, we have sets $S_1, \dots S_{n_2}$ such that

$$Y_i^{'}|Z, \tau_2^2 \sim N\left(\frac{1}{|S_i|}\sum_{j \in S_i} z_j, \tau_2^2\right),$$

where τ_1^2, τ_2^2 are the white noise variance parameters for Y and $Y^{'}$ respectively.

With minor changes to our proposed algorithm, we are able to include this data and sample from the posterior distribution for Z. In this case, when sampling Z using the (approximate) MALA algorithm, another vector-valued term must be added to the direction accordingly, the l^{th} element of which may be written as

$$\bigtriangledown \log\left(P(Y^{'}|Z, \tau_2^2)\right)_l = \frac{1}{|S_i|\tau_2^2}\left(Y_i^{'} - \frac{1}{|S_i|\tau^2}\sum_{j \in S_i} z_j\right),$$

where $l \in S_i$. By including satellite data in this way we may be able to improve spatial air pollution predictions, not only through having an extra data source, but by making the on-the-ground monitors more useful by better understanding the sampling process by which they have been selected. Similarly, if, for example, the 'on the ground' monitors were positioned preferentially in response to satellite data, knowledge about the resolution of this data may be incorporated (as in Chapter 6) in the choice of a (possibly known) discretisation.

In some cases there may be evidence to suggest that there is direct dependence between the data from one source, or taken at one time point, and the choice of sampling locations another (e.g. a preliminary survey, or a time series of surveys). In such a situation it may be beneficial to adjust the way in which preference is modelled accordingly: if a time series of spatial surfaces is being estimated, then the utility function for the 'current survey' should depend on the values of the spatial surface at the time point corresponding to an earlier survey. This notion of using time-series data to estimate levels of preference is explored in the work of Shaddick and Zidek (2014) and Watson et al. (2019), in which the location or operationality of monitors at time t is dependent on the values measured at time $t-1$. It would be straightforward to employ utility-based methods within this paradigm: by doing this we may not only get improved estimation of preference, as we would have certain knowledge of the measurements of the underlying process on which the choice of monitor locations was based, but

the inclusion of space-filling terms may be helpful if the experimenter had preferences relating to a maximum level of sparsity in an area, even if the monitors were giving low values.

7.2 Galicia lead data

We consider the widely discussed (Fernández et al. (2005), Diggle et al. (2010), Dinsdale and Salibian-Barrera (2019)) Galicia lead bryomonitoring data. This is comprised of measurements of lead concentrations found in samples of moss collected from various sites across Galicia, in North Western Spain. The data were collected in two rounds: one set in 1997, in which the monitoring sites are more heavily clustered in the north of the region, and another in the year 2000, in which the monitors are arranged in a grid-like structure over the whole region. Fernández et al. (2005) describe how the 1997 sampling sites were selected to be in places with a high concentration gradient. This data is available from the 'PrevMap' R package of Giorgi and Diggle (2017).

Initial exploration indicates that the lead concentrations do not remain constant over the three intervening years. Around this time leaded petrol (a major source of environmental lead pollution), was being phased out in the European Union, before a ban on its sale on the 1st of Jan 2000, with Spain being granted an extension until 2002 (United Nations Economic Commission for Europe (2003)). In reflection of this we, (as in Dinsdale and Salibian-Barrera (2019)) allow different means θ_{1997}, θ_{2000} for the 1997 and 2000 Gaussian random fields, predicted from the log measurements, while assuming shared values of φ, σ^2 and τ^2.

We predict the 1997 and 2000 log concentrations over a regular 45×45 grid with the grid points falling outside the Galicia boundary removed. For comparison, we fit several model and data set combinations, allowing for preference vs. assuming no preference in the 1997 data, and using only the 1997 data, only the 2000 data, and both sets combined. We denote the values of Z at the regular grid points pertaining to the 1997 and 2000 surveys by Z_{1997} and Z_{2000} respectively, and likewise the discretised-space designs D by D_{1997} and D_{2000} .

The models we fit, when predicting using information from both data sets are largely the same as those fitted in Chapter 5 to the Scottish field ammonia data, but with some differences as a result of using two sources of data. In particular, when data from both years are used, we have the following full conditional distribution for the regular-grid point realisations of the Gaussian random fields:

$$Z_{1997} = Z_{2000} - (\theta_{2000} - \theta_{1997}),$$

and

$$P(Z_{2000}|X_{1997}, X_{2000}, \theta_{2000}, \theta_{1997}, \varphi, \sigma^2, Y_{1997}, Y_{2000}, D_{1997}) = P(Z_{2000}|X_{1997}, X_{2000}, \theta_{2000}, \theta_{1997}, \varphi, \sigma^2)\frac{U(Z_{1997}, D_{1997}, \alpha, \beta)}{K(Z_{1997}, \alpha, \beta)},$$

where $P(Z_{2000}|X_{1997}, X_{2000}, \theta_{2000}, \theta_{1997}, \varphi, \sigma^2)$ is a multivariate normal distribution, conditioned on the values of $X_{2000} - \theta_{2000}$ and $X_{1997} - \theta_{1997}$. Other prior distributions are as in Chapter 5, with hyperparameters $a = 2$, $b = 5$, $c = 3$, $d = 0.2$, $k_{1997} = k_{2000} = 1$, $l_{1997} = l_{2000} = 5$, $h = 20$, $f = 0.11$, $g = 2.1$

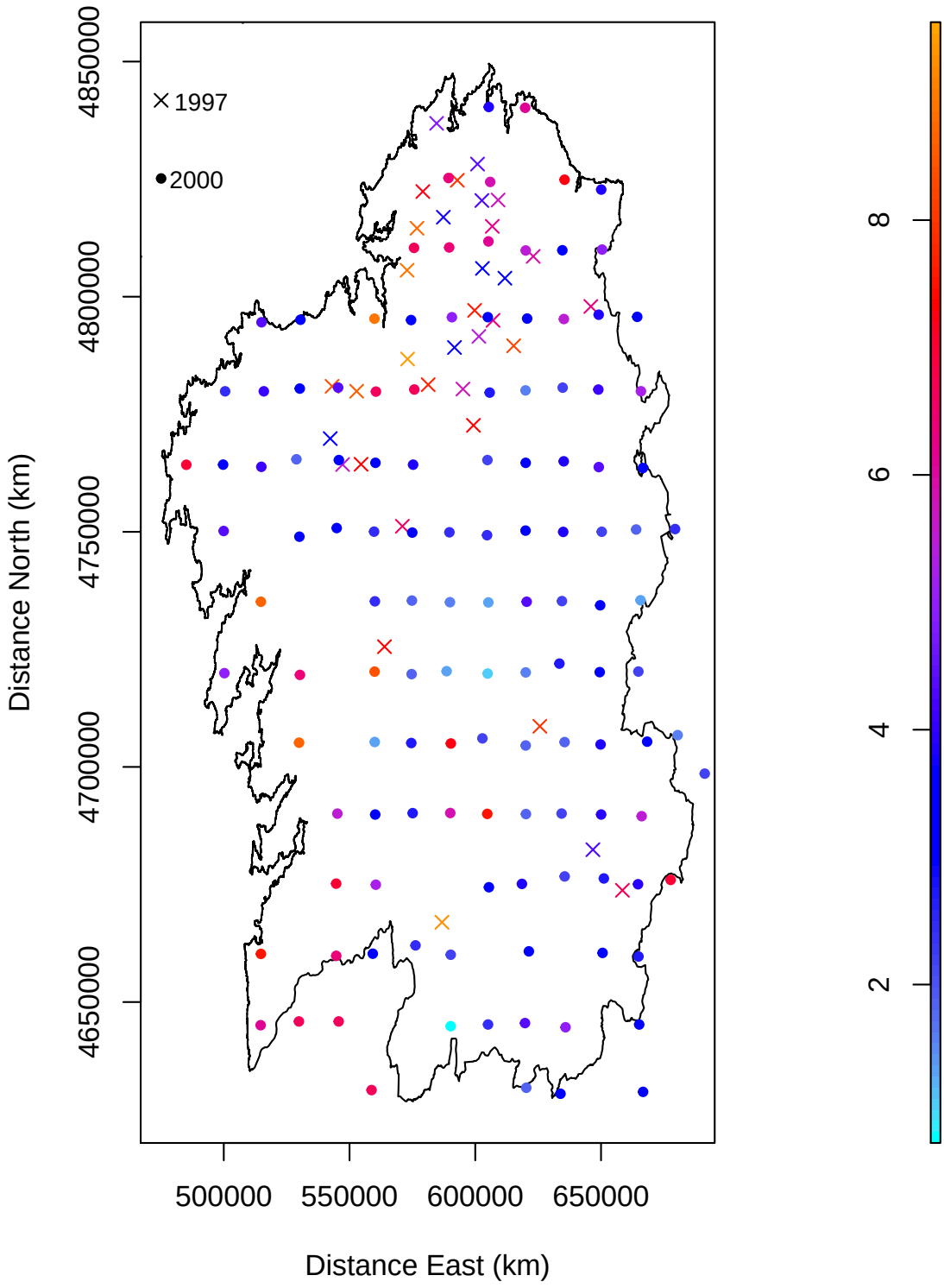


Figure 7.2: Sampling locations in the 1997 and 2000 surveys

along with a zero mean Gaussian prior distribution with variance 100 for β. We assume an exponential covariance function. The distances in kilometres from an arbitrary origin are scaled by 1×10^{-5}. We fit the model, as usual via MCMC with 10000 iterations, with the first 5000 discarded as burn-in. Z updates were made with approximate MALA proposals (as in Section 5.1.1) and, where necessary, a chain of 1000 design samples was used for each normalising constant estimation, with a burn-in of 500.

We predict the lead concentrations across the whole region using the following four models.

1. (M1) Just the 1997 data: preference allowed for via combination utility (3.8). Predictions shown are for 1997.

2. (M2) Just the 1997 data: no preference, i.e. uniform sampling assumed. Predictions shown are for 1997.

3. (M3) Just the 2000 data: no preference, i.e. uniform sampling assumed. Predictions shown are for 2000.

4. (M4) Both the 1997 data and the 2000 data: preference allowed in the 1997 samples via combination utility (3.8), while uniform sampling is assumed for the 2000 samples. Predictions shown are for 2000.

Results:

Estimated parameter values, and values of Z are shown in Table 7.1 and Figure 7.3. Parameter density and trace plots can be seen in Figures 7.4 to 7.7. For the 1997 data alone clear negative preference has been identified with $\hat{\alpha} = -2.97$, increasing the size of the areas where high values are predicted, where there are very few sampling sites, such as in the South Eastern section, in which the Galician Massif mountain range is located, when compared with the model in which grid like sampling is exhibited. When the 2000, non-preferential data is also included the strength of preference detected in the 1997 data is much weakened, to the extent that it is not conclusively indicative of a preference for negative values, with twenty percent of sampled values of α above zero. This seems, possibly, a more sensible state of affairs: it would seem odd for there to be higher levels of lead pollution in the rural south eastern areas. Comparing these results with the 2000 data only model, there is not a great deal of difference, though more lower values have been predicted in the more highly sampled north of the region. There has not been, in either of the two preference-assumed models, any substantial space-filling detected. Clearly there are some identifiability issues between the covariance parameters, however, this does not appear to have had much of an effect on the predictions of the surface Z. A possible future exploration of this data might use a utility function which, instead of rewarding either high or low values, rewarded the sampling of both extrema, which may fit better with the intentions of the experimenter of sampling from places with a high concentration gradient, as presumed by Fernández et al. (2005).

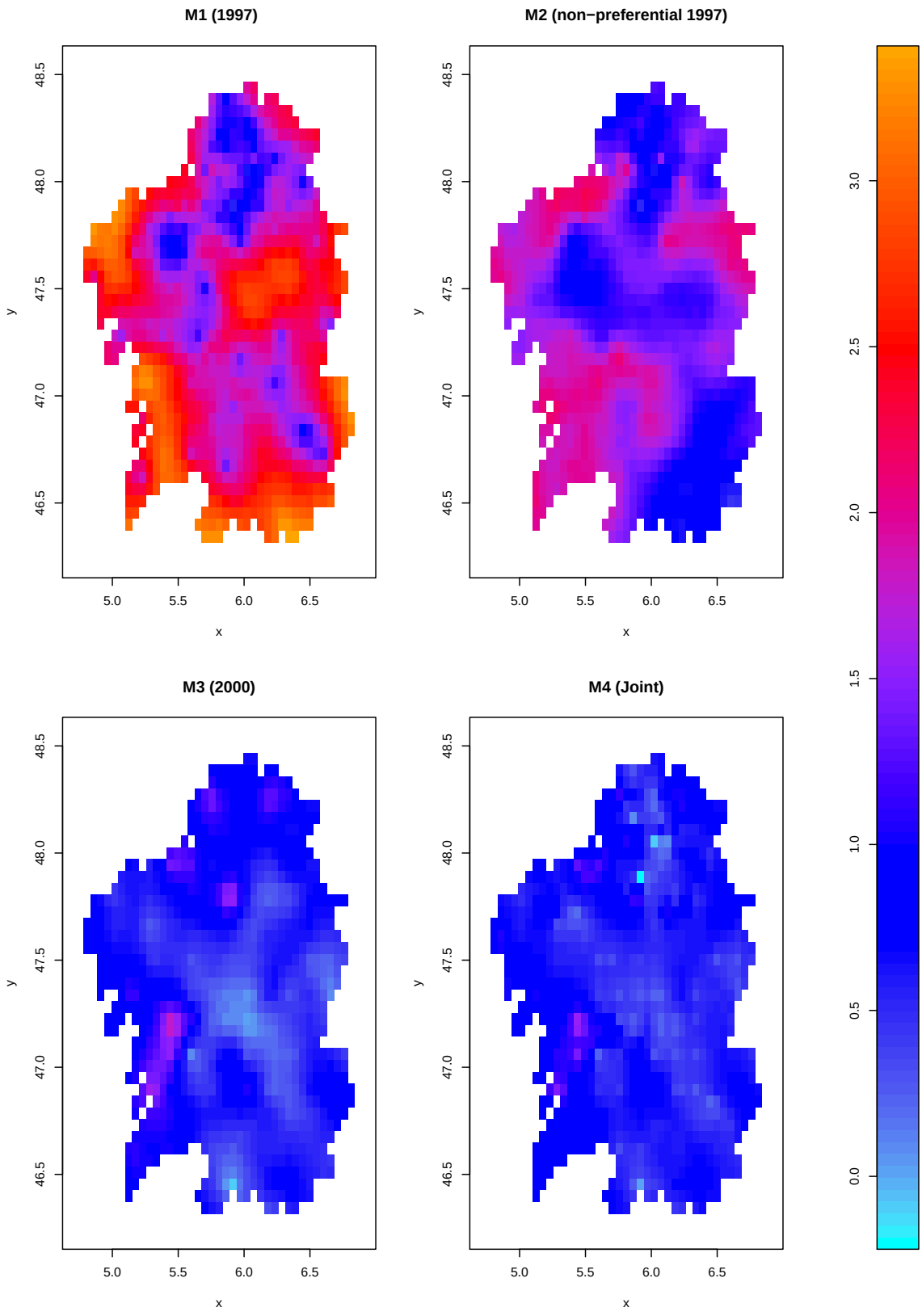


Figure 7.3: Gaussian random field estimates for the four Galicia models.

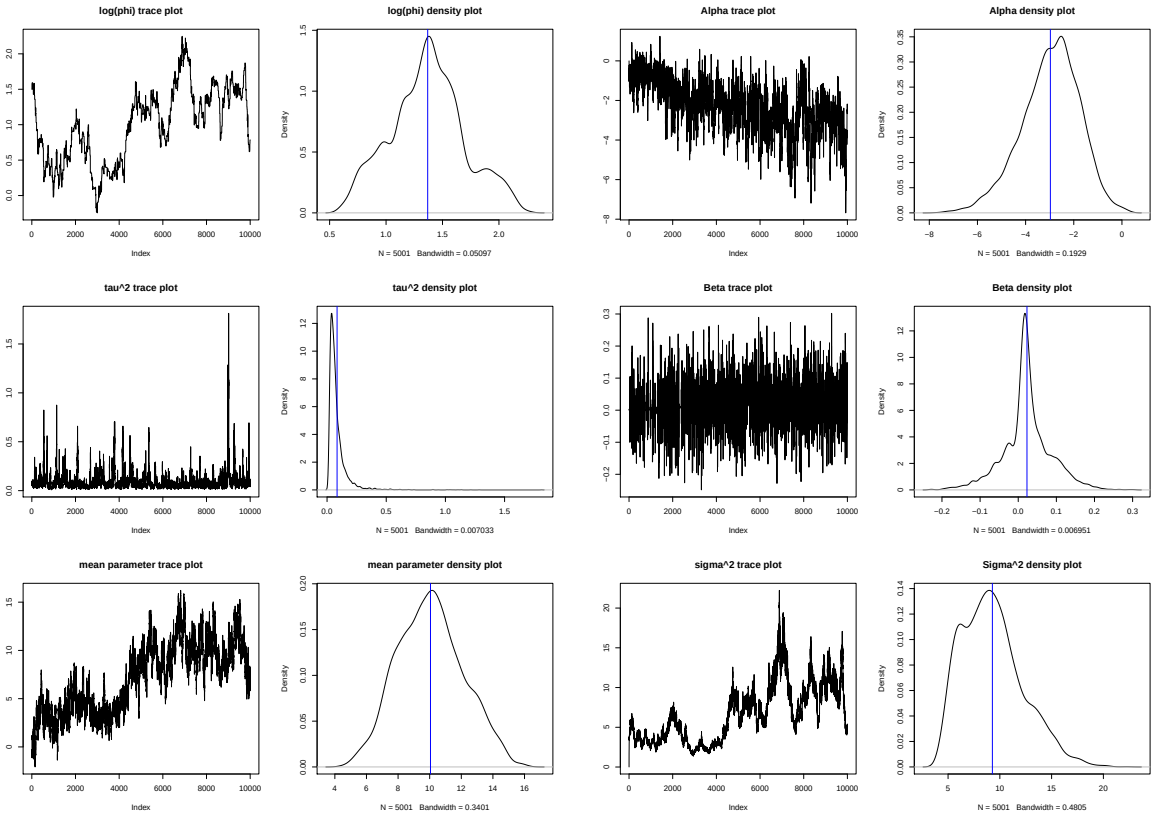


Figure 7.4: Galicia example: trace plots for model (M1).

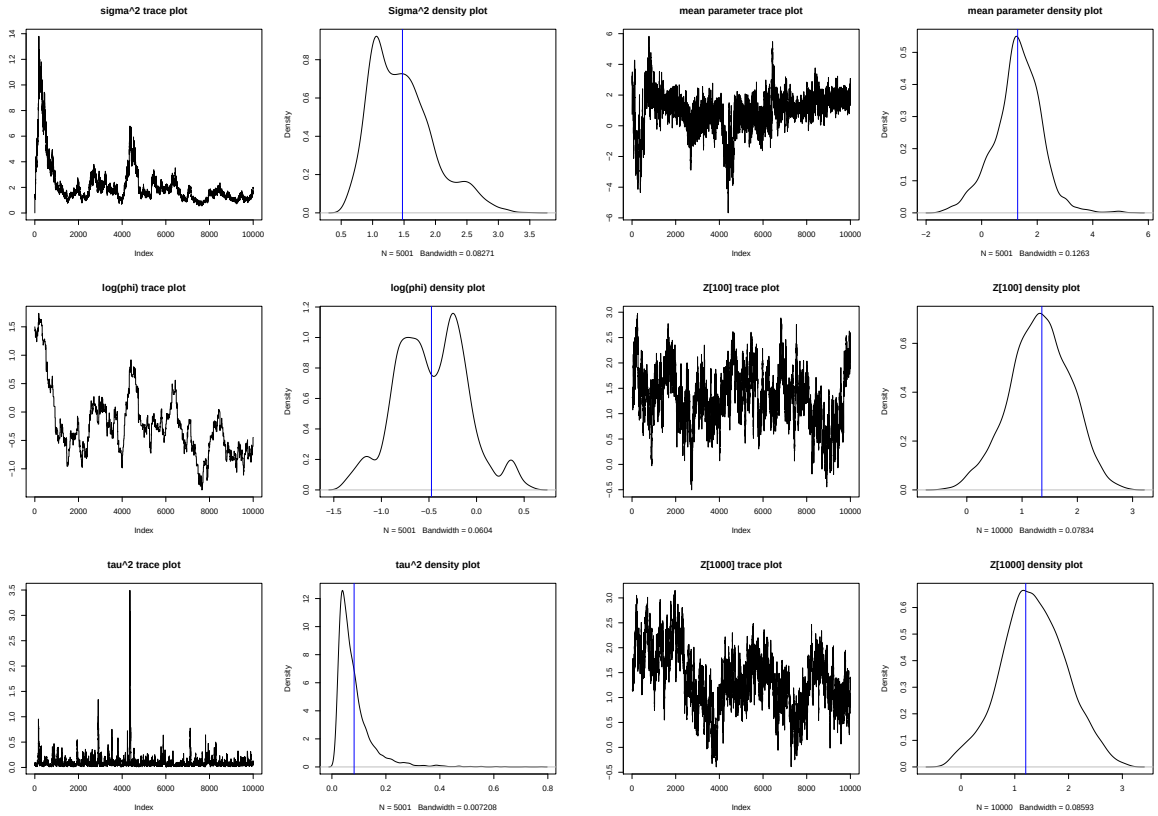


Figure 7.5: Galicia example: trace plots for model (M2).

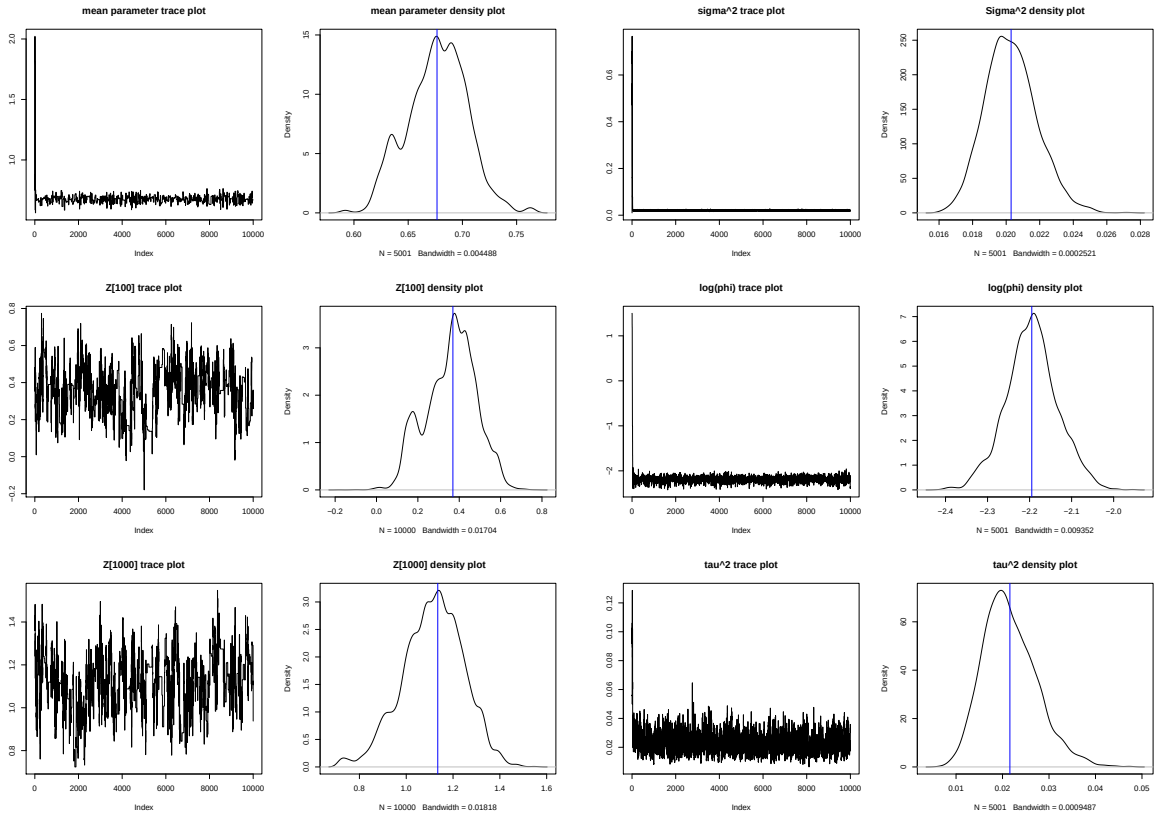


Figure 7.6: Galicia example: trace plots for model (M3).

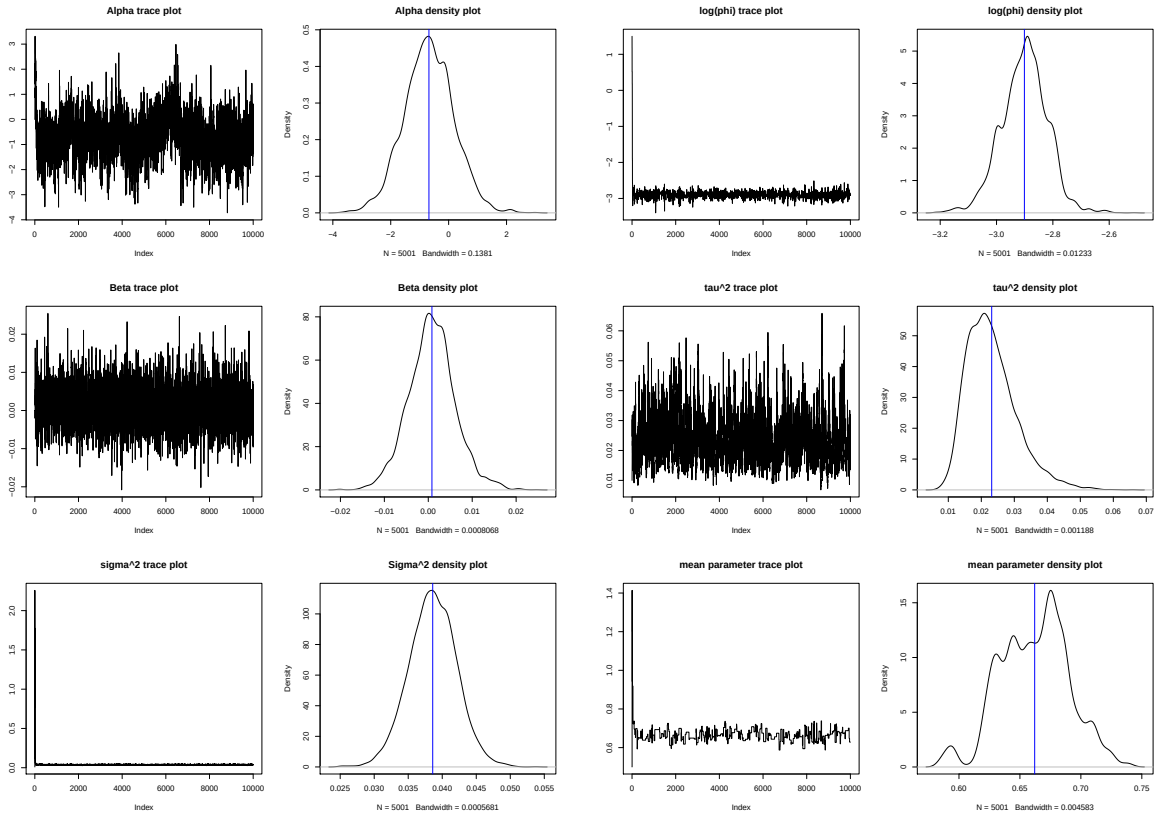


Figure 7.7: Galicia example: trace plots for model (M4).

Parameter	M1 mean (s.d.)	M2 mean (s.d.)	M3 mean (s.d.)	M4 mean (s.d.)
$\hat{\alpha}$	-2.97 (1.20)	NA	NA	-0.681 (0.861)
$\hat{\beta}$	0.0229 (0.0637)	NA	NA	0.0800 (0.530)
$\hat{\log(\varphi)}$	1.37 (0.329)	-0.474 (0.369)	-2.19 (0.0618)	-2.90 (0.0815)
$\hat{\sigma^2}$	9.28 (2.93)	1.47 (0.505)	0.0203 (0.150)	0.0386 (0.35)
$\hat{\tau^2}$	0.0847 (0.0955)	0.0822 (0.0682)	0.0216 (0.580)	0.0232 (0.760)
$\hat{\theta}_{1997}$	10.10 (2.08)	1.30 (0.86)	NA	1.47 (0.0266)
$\hat{\theta}_{2000}$	NA	NA	0.6770 (0.0274)	0.662 (0.028)

Table 7.1: Parameter estimates for the Galicia example.

Chapter 8

Conclusions and future extensions

In this thesis we have demonstrated that ignoring the sampling process of preferentially-sampled spatial data leads to inaccurate inferences, often resulting in overestimation of a process of interest in less frequently sampled areas. In particular, we have shown that existing methods, which assume the independence of sampling locations, do not sufficiently account for preference where there has been a combined preference for high values and space-filling. This is crucial, as the independence assumption indirectly implies that the 'default' sampling design would follow a uniform distribution, which is not, in many cases, realistic. Where an experimenter has a preference for spreading monitors evenly, a greater level of preference for sampling in high (or low) valued areas is required to overcome this and allow for the formation of clusters of monitors.

In order to account for a more realistic, diverse range of preferences, we have presented a general framework by which preferences may be described via a utility function, together with methods for fitting the corresponding joint models for the preference and underlying surfaces.

In terms of specific utilities, we have focused predominantly on those by which a preference for high values may be balanced with a preference for good space-filling of a region. As such, we have considered how 'good space-filling' may be quantified, and selected functions which allow for a this balancing of preferences using the mean nearest neighbour distances for both potential sampling locations, and the monitors themselves. Further extensions to this work could include the exploration of different kinds of space-filling utilities, along a road, for example, where a non-Euclidean 'along-road distance' may be employed. This would be useful for monitors that are solely positioned along a road, along which we wish to predict air pollution. Another future direction may involve the investigation of utilities with a temporal component, for example, preferences may change over time: say the quantity being monitored fell to less hazardous levels, and so a preference for detecting high values became less urgent, while a being able to map the quantity over a region was still of some interest: here we might use an α parameter which depended on time. Similarly, we might consider the case in which we have utilities based on previous monitoring networks, with higher utilities assigned to monitoring networks that share monitoring sites with existing networks. This would bring utility-based modelling of preferential sampling into a spatio-temporal setting. Investigation is already active in this area, for example, Shaddick and Zidek (2014) show how across the UK Black Smoke monitoring network over many years, preferential sampling was at work as a monitoring site was more likely to be closed down if

it was recording lower values. It is easy to see how space-filling and dependence of sampling locations may be of relevance in this area: might a monitoring station be less likely to be shut down, despite recording very low values, if it is the last one in a certain region? Clearly in such a setting adjustments must be made to the model-fitting algorithm, for example, the approximate MALA proposal we have described requires that a utility function be differentiable in terms of the process of interest Z. Another extension could be to include terms in the utility function which depend on distances to any known point sources of pollution: there are feasibly situations in which this could lead to more accurate modelling of preference as the preferential positioning of monitors in certain areas may be due to this distance, rather than directly upon assumptions about pollution levels in that area. It is not immediately clear what effect this might have on the estimation of the strength of preference parameter α, (if this distance did not account for all the preference, say), so this could be investigated further.

The application of these methods are not without computational challenges, and we have presented methods to overcome these. This has included presenting methods for sampling from design distributions to estimate intractable normalising constants, the details of the application of an approximate MALA proposal for more efficient sampling of the process of interest, and the definition of a class of 'permutation-invariant' utility functions which display properties that allow for more efficient normalising constant estimation.

We have also seen, via exploration of the Galicia lead pollution data and simulated data sets, the need for careful consideration of the suitability of the available data when modelling preference, as there are situations in which the level of preference may be difficult to estimate well. More investigation is needed here to determine whether a sample provides sufficient information to model preference effectively and robustly, or whether other sources of data must be brought in in order to make meaningful inferences. This would be dependent on sample size, monitor density (i.e. there is a need for at least some points in the lower, more sparsely sampled area in order for the strength of preference to be estimated) and covariance structure, or measurement of how sensitive the preference is to single monitoring sites.

Conversely, we have had a glimpse of the possible gains to be achieved when the level of preferential sampling associated with a particular data set is estimated using multiple sources of data. It would be interesting to apply this principle to a mixture of satellite and ground-level air quality monitoring data.

We saw in Chapter 6 that, where a multinomial utility used, the deviance information criteria may be used effectively to select an appropriate discretisation by which the design space may be defined. In this chapter we also explored the use of a random discretisation of the region of interest, but found that, in practice, this yields little benefit. In one sense this is a helpful indication that in many cases the assumption that we may let the utility function depend on the process of interest via the points at which we seek to predict it is valid. Nonetheless, in some cases there might be reason to explore the concept of different discretisations further, where there are strong indications of a partition dependent on geographical features. In this case it may be helpful to somehow define a prior distribution for the partition based on these features.

Another future extension could involve extending some of the methods explored to health data, both in terms of determining the effects of preferential sampling of health-related environmental quantities on health outcomes, and to situations in which preference may be dependent on measurable health-related quantities, as the concept of utility-based modelling of preference can also be relevant to observational

studies, and any situation in which the choice of sampling units is not made by the person doing the analysis. Likewise, the data-set combining methods may be useful for disease prevalence mapping for diseases with a large number of asymptomatic cases, in situations in which the disease may be detected from both symptomatic cases and environmental surveillance. Modelling preference from these two sources may lead to better understanding of the spatial distribution of the disease, and ultimately, better directed responses.

Bibliography

Alquier, P., Friel, N., Everitt, R., and Boland, A. (2016). Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels. *Statistics and Computing*, 26(1-2):29–47.

Andrieu, C. and Livingstone, S. (2019). Peskun-Tierney ordering for Markov chain and process Monte Carlo: beyond the reversible scenario. *arXiv preprint arXiv:1906.06197*.

Christensen, O. F., Møller, J., and Waagepetersen, R. (2000). Analysis of spatial data using generalized linear mixed models and langevin-type markov chain monte carlo.

Christensen, O. F., Roberts, G. O., and Sköld, M. (2006). Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 15(1):1–17.

Conn, P. B., Thorson, J. T., and Johnson, D. S. (2017). Confronting preferential sampling when analysing population distributions: diagnosis and model-based triage. *Methods in Ecology and Evolution*, 8(11):1535–1546.

Cowan, N.J.; Levy, P. S. U. (2019). Nitrous Oxide fluxes and associated soil measurements from a mixed livestock farm in central Scotland (2012-2013). NERC Environmental Information Data Centre. `https://doi.org/10.5285/54edbdcf-086e-40a7-b2cc-c1e4fcbfbbbc`.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

da Silva Ferreira, G. (2020). Geostatistics under preferential sampling in the presence of local repulsion effects. *Environmental and Ecological Statistics*, pages 1–22.

da Silva Ferreira, G. and Gamerman, D. (2015). Optimal design in geostatistics under preferential sampling. *Bayesian Analysis*, 10(3):711–735.

Diaconis, P., Holmes, S., and Neal, R. M. (2000). Analysis of a nonreversible Markov chain sampler. *Annals of Applied Probability*, pages 726–752.

Diggle, P. J., Menezes, R., and Su, T.-l. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2):191–232.

Dinsdale, D. and Salibian-Barrera, M. (2019). Methods for preferential sampling in geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(1):181–198.

European Space Agency (2019). Nitrogen dioxide pollution mapped. `www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-5P/Nitrogen_dioxide_pollution_mapped`.

Fernández, J., Real, C., Couto, J., Aboal, J., and Carballeira, A. (2005). The effect of sampling design on extensive bryomonitoring surveys of air pollution. *Science of the Total Environment*, 337(1-3):11–21.

Flegal, J. M. and Jones, G. L. (2010). Batch means and spectral variance estimators in markov chain monte carlo. *The Annals of Statistics*, 38(2):1034–1070.

Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M. (2010). *Handbook of spatial statistics*. CRC press.

Gelfand, A. E., Sahu, S. K., and Holland, D. M. (2012). On the effect of preferential sampling in spatial prediction. *Environmetrics*, 23(7):565–578.

Gerber, H. U. and Pafum, G. (1998). Utility functions: from risk theory to finance. *North American Actuarial Journal*, 2(3):74–91.

Geyer, C. J. (1994). Estimating normalizing constants and reweighting mixtures. *Technical Report, School of Statistics, University of Minnesota, Minneapolis 568*.

Giorgi, E. and Diggle, P. (2017). PrevMap: An R package for prevalence mapping. *Journal of Statistical Software*, 78.

Grisotto, L., Consonni, D., Cecconi, L., Catelan, D., Lagazio, C., Bertazzi, P. A., Baccini, M., and Biggeri, A. (2016). Geostatistical integration and uncertainty in pollutant concentration surface under preferential sampling. *Geospatial Health*, 11(1).

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Jambeck, J. R. and Johnsen, K. (2015). Citizen-based litter and marine debris data collection and mapping. *Computing in Science & Engineering*, 17(4):20–26.

Kim, H.-M., Mallick, B. K., and Holmes, C. (2005). Analyzing nonstationary spatial data using piecewise Gaussian processes. *Journal of the American Statistical Association*, 100(470):653–668.

Krige, D. (1951). A statistical approach to some mine valuations and allied problems at the witwatersrand. *Master's thesis of the University of Witwatersrand*.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Lee, A., Szpiro, A., Kim, S., and Sheppard, L. (2015). Impact of preferential sampling on exposure prediction and health effect inference in the context of air pollution epidemiology. *Environmetrics*, 26(4):255–267.

Liang, F. and Jin, I.-H. (2013). A Monte Carlo Metropolis-Hastings algorithm for sampling from distributions with intractable normalizing constants. *Neural computation*, 25(8):2199–2234.

Lindley, D. V. (1975). *Making decisions*. Wiley Interscience, 3rd edition.

Michalcová, D., Lvončík, S., Chytrỳ, M., and Hájek, O. (2011). Bias in vegetation databases? a comparison of stratified-random and preferential sampling. *Journal of Vegetation Science*, 22(2):281–291.

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). Log Gaussian Cox processes. *Scandinavian journal of statistics*, 25(3):451–482.

Müller, P. (2005). Simulation based optimal design. *Handbook of Statistics*, 25:509–518.

Nychka, D., Yang, Q., and Royle, J. A. (1997). Constructing spatial designs using regression subset selection. *Statistics for the Environment*, 3:13.

Pati, D., Reich, B. J., and Dunson, D. B. (2011). Bayesian geostatistical modelling with informative sampling locations. *Biometrika*, 98(1):35–48.

Pillai, N. S., Stuart, A. M., and Thiéry, A. H. (2012). Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions. *The Annals of Applied Probability*, 22(6):2320–2356.

Pope, C. A., Gosling, J. P., Barber, S., Johnson, J. S., Yamaguchi, T., Feingold, G., and Blackwell, P. G. (2019). Gaussian process modeling of heterogeneity and discontinuities using Voronoi tessellations. *Technometrics*, pages 1–20.

Pronzato, L. and Müller, W. G. (2012). Design of computer experiments: space filling and beyond. *Statistics and Computing*, 22(3):681–701.

Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268.

Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363.

Scott, A. J. and Wild, C. J. (2011). Fitting regression models with response-biased samples. *Canadian Journal of Statistics*, 39(3):519–536.

Shaddick, G. and Zidek, J. (2014). A case study in preferential sampling: Long term monitoring of air pollution in the UK. *Spatial Statistics*, 9:51–65.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series B (Statistical Methodology)*, 64(4):583–639.

Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.2.

United Nations Economic Commission for Europe (2003). Progress report on the implementation of the pan-European strategy to phase out leaded petrol. Submitted by the Ministry of Environment of Denmark through the Ad Hoc Working Group of Senior Officials, for the fifth ministerial conference: Environment for Europe, Kiev 21-23 May 2003.

Vats, D., Flegal, J. M., and Jones, G. L. (2019). Multivariate output analysis for Markov chain Monte Carlo. *Biometrika*, 106(2):321–337.

Watson, J., Zidek, J. V., Shaddick, G., et al. (2019). A general theory for preferential sampling in environmental networks. *Annals of Applied Statistics*, 13(4):2662–2700.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261.

Zidek, J. V., Shaddick, G., Taylor, C. G., et al. (2014). Reducing estimation bias in adaptively changing monitoring networks with preferential site selection. *Annals of Applied Statistics*, 8(3):1640–1670.